Multi-Resource Parallel Query Scheduling and Optimization

Minos N. Garofalakis*
Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974
minos@research.bell-labs.com

Yannis E. Ioannidis
Department of Informatics
University of Athens
Hellas (Greece)
yannis@di.uoa.gr

Abstract

Scheduling query execution plans is a particularly complex problem in shared-nothing parallel systems, where each site consists of a collection of local time-shared (e.g., CPU(s) or disk(s)) and space-shared (e.g., memory) resources and communicates with remote sites by message-passing. Earlier work on parallel query scheduling employs either (a) one-dimensional models of parallel task scheduling, effectively ignoring the potential benefits of resource sharing, or (b) models of globally accessible resource units, which are appropriate only for shared-memory architectures, since they cannot capture the affinity of system resources to sites. In this paper, we develop a general approach capturing the full complexity of scheduling distributed, multi-dimensional resource units for all forms of parallelism within and across queries and operators. We present a level-based list scheduling heuristic algorithm for independent query tasks (i.e., physical operator pipelines) that is provably near-optimal for given degrees of partitioned parallelism (with a worst-case performance ratio that depends on the number of time-shared and space-shared resources per site and the granularity of the clones). We also propose extensions to handle blocking constraints in logical operator (e.g., hash-join) pipelines and bushy query plans as well as on-line task arrivals (e.g., in a dynamic or multi-query execution environment). Experiments with our scheduling algorithms implemented on top of a detailed simulation model verify their effectiveness compared to existing approaches in a realistic setting. Based on our analytical and experimental results, we revisit the open problem of designing efficient cost models for parallel query optimization and propose a solution that captures all the important parameters of parallel execution.

*Part of this work was done while the first author was a graduate student at the University of Wisconsin–Madison.

1 Introduction

Parallelism has been recognized as a powerful and cost-effective means of handling the projected increases in data size and query complexity in future database applications. Among all proposals, the shared-nothing [Sto86, DGS+90, DG92] and, recently, the more general hierarchical (or, hybrid) [Gra93, BFV96, NZT96] multiprocessor architectures have emerged as the most scalable to support very large database management. In these systems, each *site* consists of its own set of local resources (such as, memory, CPU(s), and disk(s)) and communicates with other sites only by message-passing over a high-speed interconnect. Despite the popularity of these architectures, the development of effective and efficient query processing and optimization techniques to exploit their full potential still remains an issue of concern [GHK92, Val93, HFV96].

Perhaps the major difference between parallel query optimization and its well-understood centralized counterpart lies in the choice of *response time* as a more appropriate optimization metric for parallel database systems. Ideally, this means that a scheduling algorithm should be run on every candidate query execution plan so that its response time is estimated while taking parallelism into account. To avoid the significant increase in optimization cost that this implies, a *two-phase optimization* is often advocated: the optimal sequential plan is identified in a first phase using the traditional, work (i.e., total resource consumption), metric for plan comparison; a scheduling algorithm is run on that plan in a second phase to identify its optimal parallelization. Although prior work has demonstrated that this often leads to plans that are inherently sequential and, consequently, unable to exploit the available parallelism [JPS93, BFG+95, LVZ93], the technique is quite attractive due to its speed. Whether in a one-phase or a two-phase optimizer, however, it is clear that effective scheduling algorithms are still needed to intelligently *parallelize* query execution plans [HM94, CHM95, Has95].

The above observation has lead to significant research activity on the complex problem of *parallelizing* a query execution plan; that is, scheduling the plan's operators to execute on the resources of a parallel system with the objective of minimizing the response time of the query [CHM95, GW93, HM94, Has95, Hon92, HCY94, LCRY93, Sch90a]. Most of these earlier efforts, however, are based on simplifying assumptions (e.g., restricted forms of intra-query parallelism) that tend to limit their applicability.

One of the main sources of complexity of query plan scheduling is the *multi-dimensionality* of the resource needs of database queries; that is, during their execution queries typically require multiple different resources, such as memory buffers, and CPU and disk bandwidth. It is therefore important to employ an appropriate multi-dimensional cost model that can capture a query's demand for individual system resources as separate components (i.e., dimensions). The reason, of course, is that such a model introduces a range of possibilities for effectively sharing system resources among concurrent query operators, which can substantially increase the utilization of these resources and reduce the response time of the query. Moreover, system resources can be categorized into two radically different classes with respect to their mode of usage by query plan operators:

- Time-Shared (TS) (or, preemptable) resources (e.g., CPUs, disks, network interfaces), that can be sliced across concurrent operators at very low overhead [GHK92, GI96, GI97]. For a TS resource, operators specify an amount of *work* (i.e., the effective time for which the resource is used) that can be appropriately "stretched" over the operator's execution time (which, of course, depends on the level of contention at the TS resource).

- Space-Shared (SS) resources (e.g., memory buffers), whose time-sharing among concurrent operators introduces prohibitively high overheads [GHK92]. For a SS resource, operators typically specify rigid

capacity requirements that must be satisfied throughout their execution[1]. Of course, this means that the total SS requirements of concurrent operators cannot exceed the available SS resource capacity.

Example 1.1: Consider a parallel database system, with each site comprising one CPU, one disk, and M pages of main memory (i.e., three dimensions). Also, consider a relation R_1 whose tuples are stored on disk at some site of the system and a selection operator $\texttt{select}(R_1.A > 10)$, where A is a numeric attribute of R_1. Using DBMS catalog information (e.g., existence of an index on $R_1.A$) and simple cost-model equations, we can estimate the *work* (e.g., in msec) that $\texttt{select}(R_1.A > 10)$ imposes on the CPU ($\texttt{work}_{\texttt{sel}}(CPU)$) and disk ($\texttt{work}_{\texttt{sel}}(disk)$) resources of the site. Further, assuming that `select` operators use simple double buffering, the memory requirement of $\texttt{select}(R_1.A > 10)$ from the site (in pages) is simply $\texttt{mem}_{\texttt{sel}} = 2$.

Additionally, consider a `build` operator running on the same site concurrently with $\texttt{select}(R_1.A > 10)$ to build an in-memory hash table for a different relation R_2. The two concurrent operators will obviously time-share the CPU and (possibly) the disk resource at the site, so the execution times of both operators increase; that is, their `work` requirements on each TS resource are "stretched" over a longer period of time. On the other hand, the physical memory allotted to each operator is fixed throughout its execution and memory buffers should not be shared among the two clones (in order to avoid expensive paging overheads). Thus, a requirement for the two operators to run in parallel is that their total memory requirement cannot exceed M pages (the memory capacity of the site). ∎

Most previous work on parallel query scheduling has typically ignored the multi-dimensional nature of database queries. It has simplified the allocation of TS resources to a mere allocation of processors, hiding the multi-dimensionality of query operators under a scalar "time" cost metric [CHM95, GW93, HM94, HCY94, LCRY93]. This one-dimensional model of scheduling is inadequate for database operations that impose a significant load on multiple system resources. With respect to SS resources, all previous work has concentrated on simplified models, assuming that the capacity of all such resources is infinite or that they are all *globally accessible* to all tasks [GG75, ST99, NSHL95, CM96]. Clearly, such models do not account for the physical distribution of resource units or the possibilities of SS resource fragmentation. This limits the usefulness of these models to a *shared-memory* context, where all processors have equal access to memory and disks [DG92].

Our Contributions. In this paper, we address the problem of effective *optimization-time parallelization* of complex query execution plans over shared-nothing architectures. This problem is, of course, of critical importance to both the *one-phase* [LVZ93] and the *two-phase* [HS91, Hon92, HS93] approach to parallel query optimization. We propose a general framework for multi-dimensional TS and SS resource scheduling in parallel database systems and employ it to address the query plan scheduling problem in its most general form, assuming the full variety of bushy plans and schedules that incorporate *independent* and *pipelined* forms of inter-operation parallelism as well as intra-operation (i.e., *partitioned*) parallelism. Our multi-dimensional framework represents query operator costs as pairs of *work* and *demand* vectors with one dimension per TS and SS resource, respectively. The basic idea is to accurately capture the operator's

[1] Our assumption of fixed memory requirements for query operators is sufficiently realistic (especially for optimization-time query scheduling [CHM95, Has95]) and has been used in several earlier studies of resource scheduling issues in database systems (see, for example, [BFV96, BFMV00, LCRY93]). Note, however, that *memory-adaptive operators* (like hybrid-hash joins [Sha86]) can actually operate under a range of possible memory allotments, with each allotment implying different operator demands for TS resources. Extending our work to deal with such *malleable* operators is a very challenging direction for future research.

requirements of individual TS(SS) resources in the system in the corresponding work (resp., demand) vector components. Referring back to Example 1.1, the work vector for our example $\texttt{select}(R_1.A > 10)$ operator is $[\texttt{work}_{\texttt{sel}}(CPU), \texttt{work}_{\texttt{sel}}(disk)]$ and its (one-dimensional) demand vector is simply $[\texttt{mem}_{\texttt{sel}}]$. We observe that the inclusion of the SS resource dimension(s) gives rise to certain interesting tradeoffs with respect to the degree of partitioned parallelism. Smaller degrees result in reduced communication overhead and, therefore, increased total work (i.e., TS resource requirements) for the operator execution (i.e., *coarse grain* parallel executions [GW93, GGS96]). On the other hand, larger degrees of parallelism in general imply smaller SS requirements for each operator clone, thus allowing for better load-balancing opportunities and tighter schedulability conditions. The importance of such tradeoffs for parallel query processing and optimization has been stressed earlier [HFV96] and is addressed in this work.

Based on our framework, we develop a fast resource scheduling algorithm for physical operator pipelines called PipeSched that belongs to the class of *list scheduling algorithms* [Gra66]. The *response time* (or, *makespan*) of the parallel schedule produced by PipeSched is analytically shown to be within $d(1 + \frac{s}{1-\lambda})$ of the optimal schedule length for given degrees of partitioned parallelism, where d and s are the dimensionalities of the TS and SS resource vectors respectively, and λ is an upper bound on the (normalized) SS demands of any clone in the pipeline. We then extend our approach to multiple independent pipelines, using a level-based scheduling algorithm [CGJT80, TWPY92] that treats PipeSched as a subroutine within each level. The resulting algorithm, termed LevelSched, is analytically shown to be near-optimal for given degrees of operator parallelism as well. Furthermore, we propose heuristic extensions to LevelSched that allow the resulting algorithm, TreeSched to handle *blocking constraints* that can arise in logical operator pipelines and arbitrary query plans, and demonstrate how the algorithm can handle *on-line* task arrivals (e.g., in a dynamic or multi-query execution environment).

We also present the results of a performance study comparing our parallel query scheduling strategies and other existing algorithms implemented on top of a detailed simulation model [Bro94] for a shared-nothing parallel database system based on the Gamma parallel database machine [DGS+90].

Finally, we revisit the open problem of designing effective cost models that would capture the essense of parallel execution and allow traditional query optimization approaches, without explicit scheduling methods, to be effectively used for parallel query optimization. These would be applicable to both one-phase optimizers that avoid scheduling or to the first phase of two-phase optimizers. We start at the work of Ganguly et al. [GGS96], which identified two important "bulk parameters" of a parallel query execution plan, namely average work and critical path length, that are important to characterizing its expected response time. Based on our analytical and experimental results, we identify the importance of a third parameter, the *average volume* (i.e., the resource-time product) for SS resources, which captures the constraints on query execution that derive from such resources. We believe that a plan cost model that captures these three "bulk parameters" is sufficient for efficient and effective parallel query optimization.

Roadmap. The remainder of this paper is organized as follows. Section 2, presents a formulation of the query scheduling problem and an overview of our approach. Our multi-dimensional model of TS and SS resource usage and quantification of partitioned operator parallelism are developed in Section 3. In Section 4, we develop the PipeSched and LevelSched algorithms and prove constant bounds on their worst-case performance ratios. Following that, we propose heuristic extensions to deal with blocking constraints (e.g., in logical operator pipelines or bushy query plans) and discuss how our approach can handle dynamic task arrivals. Section 5 presents our experimental findings from the implementation of our query scheduling algorithms in a detailed simulation model of a parallel database system. In Section 6 we discuss the implications of our results on the design of cost models for parallel query optimizers.

Section 7 reviews related work in the areas of parallel query processing and deterministic scheduling. Finally, Section 8 concludes the paper and identifies directions for future research. Proofs of theoretical results presented in this paper can be found in Appendix A.

2 Problem Formulation

In this section, we provide a more detailed description of our target parallel execution architecture and discuss our model and terminology for parallel query execution plans. We then give a concise statement of the general query scheduling problem addressed in this paper, and provide a high-level overview of our proposed solution. Finally, we discuss the main assumptions that underlie our methodology.

2.1 System and Query Plan Model

System Architecture. We consider *shared-nothing parallel database architectures* [Sto86, DGS+90, DG92], comprising several *identical* multi-resource sites. Each system site consists of (potentially) several CPUs, several disks, and a memory shared by all CPUs. Inter-site communication is done via message-passing over a high-speed interconnect, while CPUs within a site can communicate much more effectively via shared-memory. (The presentation in this paper focuses on the shared-nothing model but, since our methodology can capture multiple resources at each system site, it is equally applicable to hierarchical architectures.)

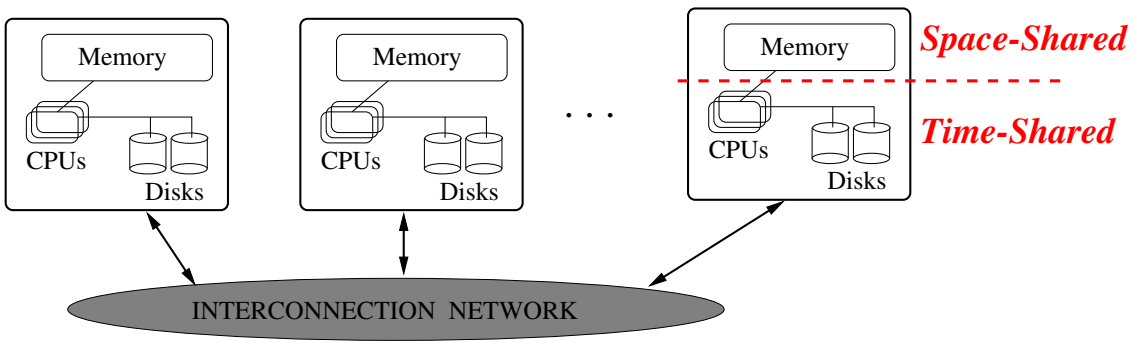

Figure 1: A shared-nothing parallel database system.

For our more abstract scheduling problem formulation, we view each site as a collection of d TS resources (capturing the site's CPUs, disks, and, possibly, other resources that are typically time-shared like network interfaces or communication processors) and s SS resources (capturing the site's memory capacity). Although memory is probably the only SS resource that comes to mind when discussing traditional query processing operators, often the distinction between time- and space-sharing depends on the needs of a particular application. As an example, the playback of a digitized video stream from disk is an operation that requires a specific fraction of disk bandwidth throughout its execution [ÖRS96, ÖRS97, GIÖ98]. Clearly, such an operation views the disk as an SS resource with a fixed bandwidth capacity, whereas traditional query processing operators do not have similar "hard real-time" requirements for data delivery and view it as a TS resource. For this reason, we have decided to address the scheduling problems for general s rather than restricting our discussion to the special case of $s = 1$ (i.e., a single dimension for memory capacity).

An obvious advantage of this general formulation is that it allows us the flexibility to "draw the line" between time- and space-sharing at any boundary, depending on factors such as application requirements or user view of resources.

Query Execution Plan Model. A *query execution plan* comprises a tree of *logical operator* nodes, like `hash-join` and `sortmerge-join` (Figure 2(a)). The shape of the plan tree can be left-deep, right-deep, or bushy. Bushy trees capture the most general space of query execution plans and are the most appealing for parallel query execution, since they offer the best opportunities for minimizing the size of intermediate results [SYT93] and for exploiting all forms of parallelism [Val93]. Thus, in this paper, we consider the full space of bushy query execution plans.

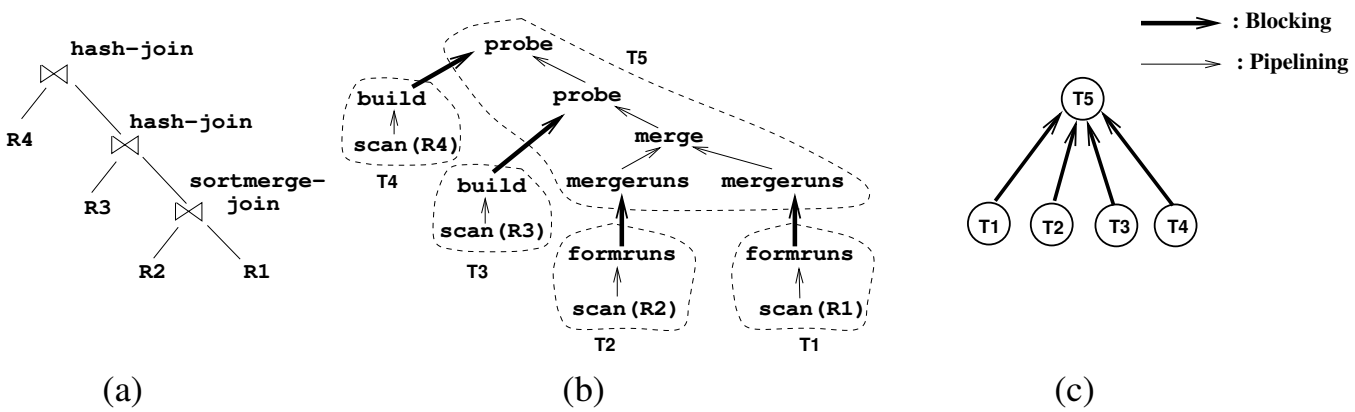

Figure 2: (a) A right-deep query execution plan. (b) The corresponding operator tree. (c) The corresponding query task tree. The thick edges in (b) indicate blocking constraints.

An *operator tree* [GHK92, Hon92, Sch90a] is created as a "macro-expansion" of a query execution plan by refining each plan node into a subtree of *physical operator* nodes, such as `scan`, `probe`, and `build` (Figure 2(b)). The operator tree representation exposes the available parallelism in the plan as well as the specific data flow and timing constraints between physical operators. In our example operator tree of Figure 2(b), edges represent the flow of data as well as the two forms of timing constraints between physical operators: *pipelining* (thin edges) and *blocking* (thick edges). Pipelining edges connect physical operators that work in parallel in "producer/consumer" mode, whereas blocking edges indicate that a parent operator cannot start execution until all its children via blocking edges have produced their full result. Furthermore, blocking constraints often imply two forms of stronger *execution dependencies* between physical operators.

1. *Disk-Materialization Dependencies.* If the child (tail) operator of a blocking edge materializes its result on disk, then the parent (head) operator (e.g., `scan`) has to be executed at the same set of sites as the child, since it needs to access the same set of disks. For example, this is the case with the `mergeruns` and `formruns` operators of Figure 2(b): each `mergeruns` needs to access the sorted runs stored on disk by its child `formruns`.

2. *Memory-Materialization Dependencies.* If child operators of blocking edges materialize their result in memory, then not only do their parent operators have to be executed at the same set of sites, but it also often needs to be guaranteed that (a) all children are executed in parallel, and (b) their parents are executed immediately afterwards as well. This is, for instance, the case with the `build` operators

of Figure 2(b), which must build their hash tables in memory in parallel, so that the corresponding `probe` parent operators, being executed immediately after them, find those tables in memory.

A *query task* is a maximal subgraph of the operator tree containing only pipelining edges. Intuitively, a query task is a fully-pipelined collection of physical operators. A *query task tree* is created from an operator tree by representing query tasks as single nodes (Figure 2(c)). Note that a single logical operator pipeline, like the right-deep execution plan of Figure 2(a), can give rise to multiple query tasks (i.e., physical operator pipelines) with various blocking constraints and execution dependencies between them.

The above tree representations clarify the definitions of the three forms of intra-query parallelism:

- *Partitioned parallelism:* A single node of the operator tree (i.e., physical operator) is executed on a set of sites by appropriately partitioning its input data set(s) across independent *operator clones*.
- *Pipelined parallelism:* The operators of a single node of the task tree (i.e., physical operator pipeline) are executed on a set of sites in a pipelined manner.
- *Independent parallelism:* Nodes of the task tree with no path between them are executed in parallel on a set of sites independent of each other (as mentioned above, they *must* be executed in parallel in certain cases). For example, in Figure 2, tasks T1-T4 may all be executed in parallel, with T3-T4 having no other choice but to be executed in parallel. Task T5 must await the completion of T1-T4 before it is executed, which should be immediately after T3-T4 complete.

The *home* of an operator is the set of sites allotted to its execution. Each operator is either *rooted*, if its home is fixed by execution dependencies (e.g., scanning the materialized result of a child operator from disk or `prob`ing the memory-resident result of a child operator), or *floating*, if the resource scheduler is free to choose its home.

2.2 Problem Statement and Overview of Approach

Consider a parallel query optimizer and a query plan tree that the optimizer has generated during the course of its exploration for the optimal plan. A *parallel execution schedule* for the query plan consists of an allocation of system resources to physical operators in the corresponding operator tree and a mapping of these operators onto system sites, such that (1) the resource needs of all operators are satisfied without violating system capacity constraints, and (2) all timing constraints and execution dependencies in the plan are respected. The general optimization-time query scheduling problem addressed in this work can be stated as follows.

Given: A query execution plan to be parallelized over a shared-nothing parallel database system, where each site comprises d TS and s SS resources.

Find: A parallel execution schedule that minimizes the response time for the input query plan.

We have devised an algorithm for parallelizing bushy query execution plans that consists of the following general steps[2]:

1. Construct the corresponding operator and task trees, and for each operator, determine its individual resource requirements using hardware parameters, DBMS statistics, and conventional optimizer cost models [SAC+79, HCY94]. (Part of this information would already be available from earlier actions of the optimizer.)

[2]The description here is necessarily abstract and does not directly address certain issues, such as the `build/probe` execution dependencies described earlier in this section. The details will be filled in later in the paper.

2. For each (floating) logical operator, determine the degree of parallelism based on the TS vs. SS resource tradeoffs mentioned earlier and discussed in detail in Section 3, taking into account any execution dependencies involving the operator (partitioned parallelism).

3. Place the tasks corresponding to the leaf nodes of the task tree in the *ready list* L of the scheduler. While L is not empty, perform the following steps:

   3.1. Determine a batch of tasks from L that can be executed concurrently (physical operator pipelines) and schedule them using a *provably near-optimal* multi-dimensional list scheduling heuristic (pipelined and independent parallelism).

   3.2. If there are tasks in the tree whose execution is enabled after Step 3.1, place them in the ready list L.

Our scheduling strategy can be readily used to handle *on-line* task arrivals (e.g., in a dynamic or multi-query execution environment).

2.3 Assumptions

Our resource scheduling model and optimality results are based on certain simplifying assumptions about the execution system. We believe that most of these assumptions are reasonable for optimization-time query scheduling, where decisions are made using a cost model that abstracts away many aspects of run-time query execution. We outline our assumptions along with a short discussion behind their rationale.

A.1 No Time-Sharing Overhead for TS Resources. Following Ganguly et al. [GHK92], we assume that slicing a preemptable resource among multiple operators introduces no additional resource costs. (Note that, a similar assumption is made in the cost model of the query optimizer for IBM's DB2 Parallel Edition [JMP97].) This is certainly realistic for TS resources with reasonably small context-switching overheads, like CPUs or communication processors. On the other hand, time-slicing a disk resource across multiple requests can lead to significant seek and latency overheads. Nevertheless, most modern disk architectures employ sequential data prefetching along multiple cache contexts, which means that they are typically able to "hide" such overheads and use the disk effectively as long as the number of concurrent requests stays below the number of cache contexts (typically 16 or higher in modern SCSI disks [Sea]). Thus, even for disk resources, we believe that our assumption is justified for reasonably high levels of concurrency. (When these levels are exceeded, a multiplicative *penalty factor* could be used to model the effects of resource contention [GHK92].)

A.2 Uniform TS Resource Usage. Following Ganguly et al. [GHK92] again, we assume that the usage of a TS resource by an operator is uniformly spread over the execution of the operator. Essentially, this implies that our scheduling model does not capture operator "hot spots" in CPU or disk usage – a reasonable assumption, we believe, for optimization-time query scheduling. Note that this assumption is clearly true *on the average* over the lifetime of the operator.

A.3 Constant SS Resource Demand. The total SS requirements of an operator are assumed to be constant and independent of its degree of parallelism; that is, distributing the execution of an operator over multiple clones does not increase its overall memory requirement. As an example, the total amount of memory required by all the clones of a `build` operator is constant and equal to the size of a hash table on the entire build relation. Further, increasing the degree of parallelism does not increase the maximum SS

demands of individual clones. These tend to be realistic assumptions for most query processing operators and all reasonable data partitioning strategies.

A.4 Dynamically Repartitioned Pipelined Inputs for Floating Operators. To compute the communication costs for floating operators and their children in an operator pipeline, we assume that the output of the children is always dynamically redistributed to serve as input to their parents. Dynamic data repartitioning is quite common, e.g., it is often required in join pipelines, when the join attributes of pipelined joins are different, the degrees of partitioned parallelism differ, or different declustering schemes must be used for load balancing. On the other hand, repartitioning is not required for parallel join strategies that try to benefit from existing data placement decisions, such as *collocated joins* (where both join inputs are already partitioned on the join attribute over the same set of sites so that no repartitioning is needed) or *directed joins* (where one of the join inputs is already partitioned on the join attribute and we only need to redistribute the tuples of the second input over the home of the first) [BFG+95]. Note, however, that for such strategies the join operators are obviously *rooted* by data placement; for example, a collocated join is always executed at the home of its operands. Thus, if the decisions of using collocated or directed joins have already been made by earlier stages of the optimizer, our scheduling framework can readily incorporate them as rooted operation costs and our assumption remains valid. If, however, these decisions are to be made during the plan scheduling step, then we believe that our scheduling framework can be extended to handle them effectively. The key observation here is that only a small, constant number of alternatives need to be explored for certain joins [BFG+95]. Consequently, at the cost of an increase in complexity, our scheduler can investigate these alternatives and select the best one.

3 Quantifying Partitioned Parallelism

In this section, we offer a formal description of our multi-dimensional model of resource usage that captures the coexistence of TS and SS resources at each system site. Accounting for both TS and SS resource dimensions raises some interesting tradeoffs with respect to the degree of partitioned parallelism for operators. We propose a quantification of these tradeoffs that allows us to derive the degree of partitioned parallelism for operators based on system parameters and given bounds on the granularity of parallel execution.

3.1 A Multi-Dimensional Model for Time- and Space-Shared Resources

For TS resources, our treatment builds on the model of preemptable resources proposed by Ganguly et al. [GHK92], which we briefly describe here. The usage of a single TS resource by an operator is modeled by two parameters, T and W, where T is the elapsed time after which the resource is freed (i.e., the response time of the operator) and W is the work measured as the effective time for which the resource is used by the operator. Intuitively, during the execution of the operator, the TS resource is kept busy only for a fraction of the time equal to W/T. Further, by Assumption (A.2), the utilization of the resource is "stretched" uniformly over T. For example, a `select` with a total CPU time requirement of 100msec that runs over a period of 1sec only uses 10% of the CPU resource over that period — the remaining 90% of the CPU can be used to accommodate other concurrent operations at no overhead (Assumption (A.1)). Thus, in conjunction with our first two assumptions, this model leads to straightforward quantification of the effects of TS resource sharing.

In this paper, we propose a model that captures the coexistence of TS and SS resources at each system site and quantifies the effects of sharing such multi-resource sites among query operators. The basic idea

is to describe an isolated operator's usage of a site comprising d TS resources and s SS resources by the triple $(T^{seq}, \overline{W}, \overline{V})$, where:

- T^{seq} is the *stand-alone*, sequential execution time of the operator at the site, i.e., its execution time assuming no concurrent operators;
- $\overline{W}$ is a d-dimensional *work vector* whose components denote the work done on individual TS resources, i.e., the effective time [GHK92, GI96, GI97] for which each resource is used by the operator; and,
- $\overline{V}$ is an s-dimensional *demand vector* whose components denote the SS resource requirements of the operator throughout its execution. For notational convenience we assume that the dimensions of $\overline{V}$ are *normalized* using the corresponding SS resource capacities of a single site.

Example 3.1: Continuing from Example 1.1, Figure 3(a) depicts our generalized view of a system site with $d = 2$ TS resources (CPU and disk) and $s = 1$ SS resource (memory). (We assume a fixed numbering of system resources for all sites; for example, dimensions 1 and 2 of $\overline{W}$ correspond to CPU and disk, respectively.) Figures 3(b,c) depict the (stand-alone) usage of a system site by our example `select` and `build` operator clones. Note that, since our `build` operator clone can fit its entire hash table in memory (Assumption (A.3)), it only makes use of the CPU resource and, therefore, its stand-alone execution time is equal to its CPU work.

The work vectors of these operators are calculated based on standard cost models of the actual process of the operand. For example, assuming that there is no index on $R_1.A$ and using $\{R_1\}$ ($|R_1|$) to denote the number of tuples (resp., disk blocks) of relation R_1, we can estimate the work vector components (in msec) of our example `select` operator using the following realistic cost model:

$$\begin{aligned} W[CPU] &= (I(init_select) + I(init_IO) \cdot |R_1| + [I(read_tuple) + I(apply_pred)] \cdot \{R_1\} + \\ &\quad I(copy_tuple) \cdot sel_{A>10}(R_1) \cdot \{R_1\} + I(terminate_select)) \,/(1000 \cdot MIPS(CPU)) \\ W[disk] &= avg_block_xfer_time(disk) \cdot |R_1| + min_seek_time(disk) \cdot \left\lceil \frac{|R_1|}{blocks_per_cylinder(disk)} \right\rceil , \end{aligned}$$

where $I(op)$ denotes the number of instructions for performing operation op, sel_p is the selectivity of predicate p, and hardware parameters (like $MIPS$ and $avg_block_xfer_time$) have the obvious interpretation. (Note that our formula for $W[disk]$ assumes that R_1 is stored sequentially on disk.) Similar formulas can also be derived for the `build` operator. ∎

The execution time T^{seq} of an operator is actually a function of the operator's individual TS resource requirements, i.e., its work vector $\overline{W}$ (sometimes emphasized by using $T^{seq}(\overline{W})$ instead of T^{seq}), and the amount of *overlap* that can be achieved between processing at different resources [GI96, GI97]. This overlap is a system parameter that depends on the hardware and software architecture of the resource sites (e.g., buffering architecture for disk I/O) as well as the algorithm implementing the operator. The operator's SS resource requirements ($\overline{V}$) depend primarily on the size of its inputs and the algorithm used to implement the operator. On the other hand, the operator's work requirements ($\overline{W}$) depend on both of these parameters as well as its SS resource allotment $\overline{V}$.

We should once again stress that our assumption of fixed SS resource demands (i.e., $\overline{V}$ component values) for operators is sufficiently realistic for optimization-time query scheduling [CHM95, Has95] and has been used in several earlier studies of resource scheduling issues in database systems (e.g. , [BFV96,

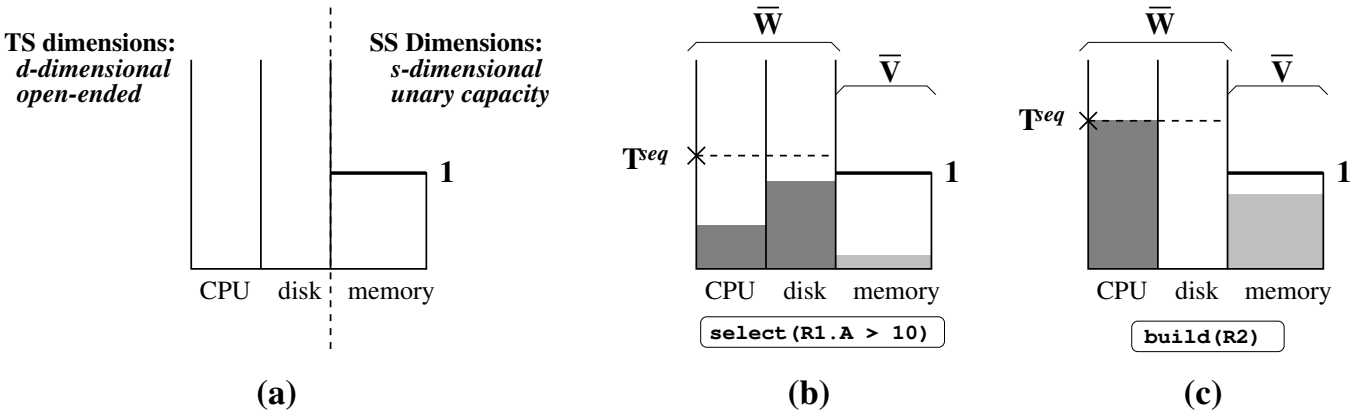

Figure 3: (a) A system site with $d = 2$ TS resources and $s = 1$ SS resource. Stand-alone usage of the site from (b) a `select` and (c) a `build` operator.

BFMV00, LCRY93]). However, most real-life query execution engines support memory-adaptive operators, whose memory requirements are typically specified as a *range* of possible memory allotments [YC93]. This flexibility adds an extra level of difficulty to our scheduling problem. It means that the scheduler also has to select, for each operator, a specific memory allotment (i.e., SS demand vector $\overline{V}$) such that the query response time is minimized over all possible $(\overline{W}, \overline{V})$ combinations. This is a very challenging generalization of our resource scheduling problem that we plan to pursue in future work.

3.2 Quantifying the Granularity of Parallel Execution

As is well known, increasing the parallelism of an operator reduces its execution time until a saturation point is reached, beyond which additional parallelism causes a speed-down, due to excessive communication startup and coordination overhead over too many sites [DGS+90]. To avoid operating beyond that point when parallelizing an operator, we want to ensure that the granules of the parallel execution are sufficiently coarse [GGS96, GI96, GI97]. However, SS resource requirements and capacity constraints introduce additional complexity to the operator parallelization problem. More specifically, in the presence of SS resources, it is not possible to concurrently execute a set of clones at a site if their total SS requirements exceed the site's capacity (in any of the s dimensions). Clearly, coarse operator clones imply that each clone has SS resource requirements that are relatively large. This means that, when restricted to coarse grain operator executions, a scheduling method can be severely limited in its ability to balance the total work across sites. Furthermore, coarse SS requests can cause fragmentation that may lead to underutilization of system resources. Thus, taking both TS and SS resources into account gives rise to interesting tradeoffs with respect to the granularity of operator clones. Our analytical results in Section 4 further demonstrate this effect.

Example 3.2: Consider the `build` operator depicted in Figure 3(c). As it stands, this `build` can only run on sites with at least $F \cdot |R_2|$ free pages of memory, where F is the standard "fudge factor" accounting for hash table overheads [Sha86]. (To simplify the notation, we assume that memory pages and disk blocks are of equal size.) Assume that we decide to partition the `build` across three clones by partitioning R_2, with the first clone getting 1/2 of R_2's tuples and the remaining two each getting 1/4 of R_2's tuples. (The

partition sizes can be estimated using statistical information on the partitioning attribute and the form of the specific hash function used.) Then, the first `build` clone can be placed at any site with $\frac{F}{2} \cdot |R_2|$ free memory pages, and the other two clones each require only $\frac{F}{4} \cdot |R_2|$ pages for their execution. This partitioning obviously gives the scheduler much more freedom to distribute the `build` work onto the system and reduce memory fragmentation at sites. On the other hand, of course, it also implies the startup and communication overheads associated with the parallel `build` execution. ∎

We now introduce some formal machinery for the purpose of quantifying the granularity of parallel operator execution and allowing us to resolve the TS vs. SS resource tradeoffs discussed above in a methodical fashion. Consider the execution of a parallel operator `op` that has been partitioned among N clones in some specific fashion. We view the granularity of that execution of `op` as a parameter that depends on the ratio $\frac{\mathcal{A}_p(\texttt{op})}{\mathcal{A}_c(\texttt{op},N)}$ and $\mathcal{V}(\texttt{op}, N)$, where:

- $\mathcal{A}_p(\texttt{op})$ denotes the total amount of work performed during the execution of `op` on a single site, when all its operands are locally resident (i.e., zero communication cost); it corresponds to the *processing area* [GW93] of `op` and is constant for all possible executions of `op`;
- $\mathcal{A}_c(\texttt{op}, N)$ denotes the total communication overhead incurred during the execution of `op` partitioned in that way; it corresponds to the *communication area* of the partitioned execution of `op` and is a non-decreasing function of N; and,
- $\mathcal{V}(\texttt{op}, N)$ denotes the maximum (normalized) SS resource requirement of any clone during the execution of `op` partitioned in that way; it corresponds to the SS *grain size* of the partitioned execution of `op` and is a non-increasing function of N.

Note that the execution of `op` with degree of partitioned parallelism equal to N is feasible only if $\mathcal{V}(\texttt{op}, N) \leq 1$; that is, the partitioning of `op` must be sufficiently fine grain for each clone to be able to maintain its SS working set at a site. The following two definitions use the notions defined above to quantify the granularity of parallel operator execution.

Definition 3.1 A parallel execution of an operator `op` with degree of partitioned parallelism equal to N is λ-*granular* if $\mathcal{V}(\texttt{op}, N) \leq \lambda$, where $\lambda \leq 1$. ∎

Definition 3.2 A parallel execution of an operator `op` with degree of partitioned parallelism equal to N is *coarse grain with parameter* f (referred to as a CG_f execution) if the communication area of the execution is no more than f times the processing area of `op`; that is, $\mathcal{A}_c(\texttt{op}, N) \leq f \cdot \mathcal{A}_p(\texttt{op})$. ∎

The granularity parameters λ and f are system-wide parameters that control the parallelization of query operators in the system. The SS granularity parameter λ essentially tries to restrict the maximum memory requirements of any operator clone. Smaller values for λ imply higher degrees of partitioned parallelism, smaller memory requirements for clones and, therefore, more scheduling freedom to "pack" these requirements onto system sites and better expected utilization of memory resources. On the other hand, the communication granularity parameter f tries to place an upper bound on the overhead of operator parallelization. Smaller values for f imply more conservative parallelizations, with fewer clones per query operator.

The λ and f parameters can obviously place conflicting requirements on the degree of operator parallelism. In our model, we give precedence to SS granularity requirements since they, in a sense, represent

"harder" constraints that can determine the schedulability of query operators. (This will become even more apparent in Section 4.) As an example, a `build` operator with a hash table larger than the size of a site's memory simply *cannot* be scheduled on a single site, even if that is what the communication granularity requirement dictates. This is more formally expressed in the following definition.

Definition 3.3 A parallel execution of an operator `op` with degree of partitioned parallelism equal to N is λ-granular CG_f, if the communication area of the execution is no more than f' times the processing area of `op`, i.e., $\mathcal{A}_c(\mathtt{op}, N) \leq f' \cdot \mathcal{A}_p(\mathtt{op})$, where f' is the minimum value greater than or equal to f such that $\mathcal{V}(\mathtt{op}, N) \leq \lambda$. ∎

Again, the intuition behind Definition 3.3 is that, because of its importance for operator schedulability, the λ-granularity restriction takes precedence over communication granularity. Thus, we may sometimes have to compromise our restrictions on communication overhead to ensure that the parallelization is in the λ-granular region. This is graphically demonstrated in Figure 4.

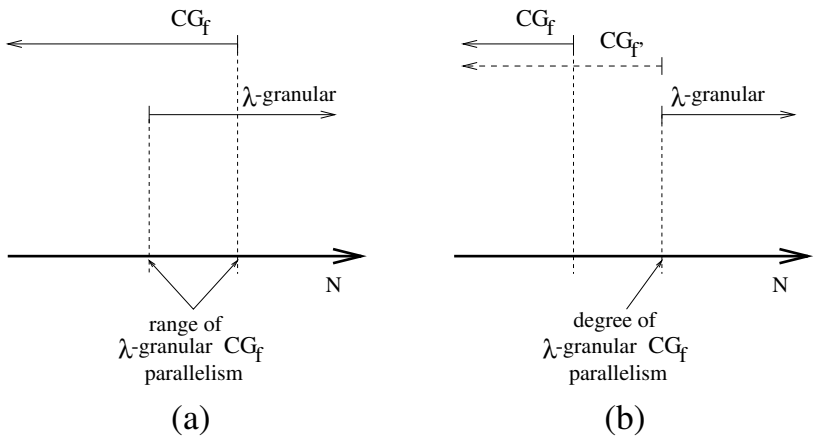

Figure 4: λ-granular CG_f execution with (a) $f = f'$, and (b) $f < f'$.

3.3 Determining the Degree of Partitioned Parallelism

We now demonstrate how our multi-dimensional resource model and our quantification of operator granularity can be employed to derive the degree of partitioned parallelism for query operators. Assuming zero communication costs, the TS and SS resource requirements of an operator `op` are described by a d-dimensional work vector $\overline{W}$ and an s-dimensional demand vector $\overline{V}$ whose components can be derived from system parameters and traditional optimizer cost models [SAC+79]. By definition, the processing area of the operator $\mathcal{A}_p(\mathtt{op})$ is simply the sum of $\overline{W}$'s components, i.e., $\mathcal{A}_p(\mathtt{op}) = \sum_{i=1}^{d} W[i]$. Similarly, the SS grain size $\mathcal{V}(\mathtt{op}, N)$ can be estimated using traditional optimizer cost models and statistics kept in the database catalogs on the distribution of attribute values [Ioa93, PI96]. Finally, we can estimate the communication area $\mathcal{A}_c(\mathtt{op}, N)$ using again some standard cost models. As such models are not ubiquitous in the database literature, we present one of them for completeness, a simple linear model of communication costs that has been adopted in previous studies of shared-nothing architectures [GMSY93, WFA92, GI96, GI97] and validated on the Gamma research prototype [DGS+90]. Specifically, if D is the total size of the operator's

input and output transferred over the interconnection network, then $\mathcal{A}_c(\mathtt{op}, N) = \alpha \cdot N + \beta \cdot D$, where α and β are architecture-specific parameters specified as follows: (1) α is the *startup cost* for each operator clone, and (2) β is the average time spent at the CPU and/or the network interface per unit of data transferred (e.g., for packaging data tuples into network messages). An experimental methodology for determining these communication cost parameters is presented by Englert et al. [EGH95] in the context of NonStop SQL/MP, a commercial shared-nothing database system from Tandem computers. Their measurements on a 4-site system show a cost of $\alpha = 0.5$ sec per clone process that needs to be created and a cost of $\beta = 1.23$ sec/MByte for data repartitioning. (They also note that α can be significantly reduced by re-using clone processes.) The following proposition is an immediate consequence of Definition 3.3.

Proposition 3.1 Assuming a linear model for clone communication costs, the maximum allowable degree of partitioned parallelism for a λ-granular CG_f execution of operator $\mathtt{op}$ is denoted by $N_{max}(\mathtt{op}, f, \lambda)$ and is determined by the formula

$$N_{max}(\mathtt{op}, f, \lambda) = \max\{\ 1\ ,\ \left\lfloor \frac{f \cdot \mathcal{A}_p(\mathtt{op}) - \beta \cdot D}{\alpha} \right\rfloor\ ,\ \min\{\ N : \mathcal{V}(\mathtt{op}, N) \leq \lambda\ \}\ \}. \quad \blacksquare$$

Proposition 3.1 provides a concise formula for determining the degree of partitioned parallelism for all floating query operators, based on our quantification of operator granularity and a concrete model for communication costs. It is important to note, however, that all our scheduling algorithms and analytical results (developed in Section 4) are *independent* of the specifics of both the processing and the communication models as well as of the quantification of communication granularity being used. In other words, our results are always valid for the given degrees of partitioned parallelism for floating operators, regardless of the exact method used to derive these degrees.

4 The Query Scheduling Algorithms

In this section, we present our multi-resource scheduling algorithms for parallel query execution plans. Our algorithms and analytical results assume that the degree of parallelism for query operators has been fixed based on system-wide granularity constraints (Section 3.2). We present the scheduling problems for parallel query plans in increasing order of complexity[3]:

1. We demonstrate how our multi-dimensional resource model can account for concurrent execution and resource sharing among independent and pipelined operator clones.

2. We present a lower bound on the optimal parallel response time for a collection of *independent* operators and propose a fast algorithm that guarantees near-optimal schedules for such operators.

3. We consider the operator scheduling problem in the presence of *pipelining* constraints and propose a novel, provably near-optimal scheduling algorithm (termed LevelSched) for independent physical operator pipelines.

4. We propose heuristic extensions that allow LevelSched to handle *blocking* constraints and execution dependencies in logical operator pipelines and bushy query plans, as well as on-line task arrivals.

We start by outlining some definitions and notational conventions that will be used in the presentation.

[3]Note that all the scheduling problems formulated in this section are $\mathcal{NP}$-hard.

4.1 Notation and Definitions

Table 1 summarizes the notation used in the development of our query scheduling algorithms with a brief description of its semantics. Detailed definitions of some of these parameters are given in the text. Additional notation will be introduced when necessary. In the rest of the paper, unless otherwise noted, all references to the term "operator" imply a physical operator.

Symbol	Semantics
P	Number of system sites
d	Number of TS resources per site
s	Number of SS resources per site
B_j	System site (i.e., "bin") j $(j = 1, \ldots, P)$
B_j^W	Set of TS work vectors scheduled at site B_j
B_j^V	Set of SS demand vectors scheduled at site B_j
$T^{site}(B_j)$	Execution time for all operator clones at site B_j
$T^{par}(\text{SCHED}, P)$	Response time for a query execution schedule SCHED on P sites
M	Number of operators to be scheduled
$\texttt{op}_i$	Operator, e.g., `scan`, `build` $(i = 1, \ldots, M)$
N_i	Degree of partitioned parallelism (number of clones) for $\texttt{op}_i$
$\overline{W}_{\texttt{op}_i}$	Work vector for $\texttt{op}_i$ (including communication costs for N_i clones)
$\overline{V}_{\texttt{op}_i}$	Demand vector for $\texttt{op}_i$
$T^{max}(\texttt{op}_i, N_i)$	Maximum execution time among the N_i clones of $\texttt{op}_i$ while alone in system
S	Set of (floating) operator clones to be scheduled
S_{C_i}	Set of (floating) clones in a physical operator pipeline C_i
S^W (S^V)	Set of work (resp., demand) vectors for all clones in S
S^{TV}	Set of *volume* (time $\times$ demand) vectors for all clones in S
$T^{max}(S)$	Maximum stand-alone execution time among all operator clones in S
$l(\overline{v}), l(S^v)$	Length of a vector $\overline{v}$ or set of vectors S^v

Table 1: Notation

The collection of physical query operators to be scheduled is denoted by $\{\texttt{op}_1, \ldots, \texttt{op}_M\}$ and N_i denotes the fixed degree of partitioned parallelism (i.e., number of clones) for operator $\texttt{op}_i$ (as determined by communication and memory granularity constraints). The work and demand vector for each operator $\texttt{op}_i$ are defined as earlier, except that communication costs for the given degree of parallelism N_i have already been factored in the operator's work requirements. More specifically:

- The demand vector $\overline{V}_{\texttt{op}_i}$ describes the total (normalized) SS resource requirements of $\texttt{op}_i$. The components of $\overline{V}_{\texttt{op}_i}$ are computed using architectural parameters and database statistics, and are independent of the degree of partitioned parallelism N_i.

- The work vector $\overline{W}_{\texttt{op}_i}$ describes the total (i.e., processing and communication) TS resource requirements of $\texttt{op}_i$, *given its degree of parallelism* N_i. Using the notions of communication and processing area defined in Section 3, the above is expressed as $\sum_{k=1}^{d} W_{\texttt{op}_i}[k] = \mathcal{A}_p(\texttt{op}_i) + \mathcal{A}_c(\texttt{op}_i, N_i)$. The individual components of $\overline{W}_{\texttt{op}_i}$ are computed using architectural parameters and database statistics, as well as the SS allotment for $\texttt{op}_i$ and a model for communication costs[4].

[4]The actual distribution of work among the vector's components is immaterial as far as our model is concerned.

Each operator $\mathtt{op}_i$ distributes its total TS and SS resource requirements $(\overline{W}_{\mathtt{op}_i}, \overline{V}_{\mathtt{op}_i})$ across N_i (work vector, demand vector) pairs corresponding to its clones. The TS and SS requirements for each such clone are determined based on the exact form of data repartitioning (e.g., range or hash) and statistical information on the partitioning attribute typically kept in the DBMS catalogs [PI96]. Note that, for the results presented in this paper, *no uniformity assumptions* are made about the distribution of data and work/demand vectors among clones; this distribution can be arbitrary, possibly highly skewed, as long as the memory granularity constraint λ is observed.

Given an operator clone with resource requirements $(\overline{W}, \overline{V})$ and a (stand-alone) execution time of $T = T^{seq}(\overline{W})$, we define the *volume vector* of the clone as the product $T \cdot \overline{V}$, i.e., the resource-time product[5] for the clone's execution [CM96, YC93]. We use S^W, S^V, and S^{TV} to denote the set of work, demand, and volume vectors (respectively) for a given set S of floating operator clones. The W, V, and TV superscripts are used in this manner throughout the paper. For example, B_j^W and B_j^V denote the sets of work and demand vectors for all clones mapped to site B_j by our scheduler.

The *length of an n-dimensional vector* $\overline{v}$ is its maximum component. The *length of a set* S^v *of n-dimensional vectors* is the maximum component in the vector sum of all the vectors in S^v. More formally,

$$l(\overline{v}) = \max_{1 \le k \le n} \{v[k]\} \quad , \quad l(S^v) = \max_{1 \le k \le n} \{ \sum_{\overline{v} \in S^v} v[k] \}.$$

The *performance ratio* of a scheduling algorithm is defined as the ratio of the response time of the schedule it generates over that of the optimal schedule. All the scheduling problems addressed in this paper are non-trivial generalizations of traditional multiprocessor scheduling [GJ79] and, thus, they are clearly $\mathcal{NP}$-hard. Given the intractability of the problems, we develop polynomial-time heuristics that are *provably near-optimal*, i.e., with a constant bound on the performance ratio.

Since the parallelization of rooted operators is pre-determined, our algorithms are only concerned with the scheduling of floating operators. Furthermore, they assume that the degree of partitioned parallelism for all floating operators has already been determined based on system-wide granularity constraints (Section 3.2); for example, as shown in Proposition 3.1.

4.2 Modeling Parallel Execution and Resource Sharing

We present a set of extensions to the (one-dimensional) cost model of a traditional DBMS based on the multi-dimensional resource usage formulation described in Section 3.1. Our extensions account for all forms of parallelism and quantify the effects of sharing TS and SS resources on the response time of a parallel execution.

4.2.1 Partitioned and Independent Parallelism

For partitioned parallelism, the work and demand vectors of an operator $\mathtt{op}_i = (\overline{W}_{\mathtt{op}_i}, \overline{V}_{\mathtt{op}_i})$ are partitioned among a set of N_i independent operator clones, where each clone executes on a single site and works on a portion of the operator's data. Given such a partitioning $\{(\overline{W}^1_{\mathtt{op}_i}, \overline{V}^1_{\mathtt{op}_i}), (\overline{W}^2_{\mathtt{op}_i}, \overline{V}^2_{\mathtt{op}_i}), \ldots, (\overline{W}^{N_i}_{\mathtt{op}_i}, \overline{V}^{N_i}_{\mathtt{op}_i})\}$, where $\sum_{k=1}^{N_i} \overline{W}^k_{\mathtt{op}_i} = \overline{W}_{\mathtt{op}_i}$ and $\sum_{k=1}^{N_i} \overline{V}^k_{\mathtt{op}_i} = \overline{V}_{\mathtt{op}_i}$, a lower bound on the parallel execution time for $\mathtt{op}_i$ is the maximum of the sequential execution times of its N_i clones; that is, the parallel execution time for

[5] The *volume* of an operator is defined as the product of the amount of resource(s) that the operator reserves during its execution and its execution time.

$\texttt{op}_i$ is always greater than or equal to

$$T^{max}(\texttt{op}_i, N_i) \;=\; \max_{1 \le k \le N_i} \{\, T^{seq}(\overline{W}^k_{\texttt{op}_i}) \,\}.$$

By our definitions of the TS and SS resource classes, it is obvious that a collection of operator clones $\{(\overline{W}_1, \overline{V}_1), (\overline{W}_2, \overline{V}_2), \ldots, (\overline{W}_m, \overline{V}_m)\}$ can be executed concurrently at some system site *only if* $l(\sum_1^m \overline{V}_i) \le 1$; that is, the total SS requirements of the clones do not exceed the capacity of the site. We call such clone collections *compatible*. The following definition formalizes the notion of a schedule for independent operators in the presence of SS resource constraints.

Definition 4.1 Given a collection of M independent operators $\{\texttt{op}_i, i = 1, \ldots, M\}$ and their respective degrees of partitioned parallelism $\{N_i, i = 1, \ldots, M\}$, a *schedule* is a partitioning of the $\sum_{i=1}^M N_i$ operator clones into a collection of compatible subsets $S_1, \ldots, S_n$ followed by a mapping of these subsets to the set of available sites. ∎

The effects of time-slicing the TS resources of a site among the clones in a compatible subset S_i can be quantified as follows. Let S_i^W denote the set of work vectors for all clones in S_i. Since all clones are running concurrently, the execution time for the clones in S_i is determined by the ability to overlap the processing of TS resource requests by different clones. Specifically, under our model of preemptable resources described in Section 3.1, the execution time for all the operator clones in S_i is defined as

$$T(S_i) \;=\; \max\{\, \max_{\overline{W} \in S_i^W} \{T^{seq}(\overline{W})\} \;,\; l(S_i^W) \,\}.$$

Intuitively, the above formula states that the execution time is determined by either the slowest executing clone in S_i ($\max_{\overline{W} \in S_i^W}\{T^{seq}(\overline{W})\}$), or the load imposed at the most heavily congested TS resource at the site ($l(S_i^W)$), whichever is greater.

Example 4.1: Consider two pairs of a `select` and a `build` operator clones, whose work vectors and stand-alone execution times are depicted in Figure 5(a,b). Assume that both clone pairs are compatible (i.e., their total memory requirement does not exceed the number of memory pages available at each site) and consider their concurrent execution at a system site. For the first `select`/`build` pair, the total work requirements can be "packed" within the response time of the `select` clone; that is, because of the limited overlap between CPU and disk processing during the `select`, the CPU resource has enough "slack" to handle the concurrent `build` (Figure 5(a)). Thus, the execution time of the first pair is simply the stand-alone response time of the slower `select` clone, i.e., $T(\{\texttt{select}, \texttt{build}\}) = T^{seq}(\overline{W}_{\texttt{select}})$. Note that, the rightmost part of Figure 5(a) depicts the utilization of the site's CPU and disk bandwidth over the period $T(\{\texttt{select}, \texttt{build}\})$, under our "uniform stretching" assumption for TS resources (Assumption (A.2)).

On the other hand, when the second `select`/`build` pair is executed concurrently, the CPU slack is not sufficient to handle the `build` within the response time of the `select` (Figure 5(b)). Thus, the CPU resource essentially becomes the bottleneck and the execution time is determined by the total CPU work, i.e., $T(\{\texttt{select}, \texttt{build}\}) = l(\{\overline{W}_{\texttt{select}}, \overline{W}_{\texttt{build}}\}) = W_{\texttt{select}}[CPU] + W_{\texttt{build}}[CPU]$. Again, the rightmost part of Figure 5(b) depicts the "stretching" of the two operators' TS resource requirements over their execution period. Clearly, the bottleneck resource (i.e., CPU) is 100% utilized. ∎

Let SCHED denote any given schedule for independent operators (as described in Definition 4.1) and let $S(B_j)$ denote the collection of compatible clone subsets mapped to system site $B_j, 1 \le j \le P$ under

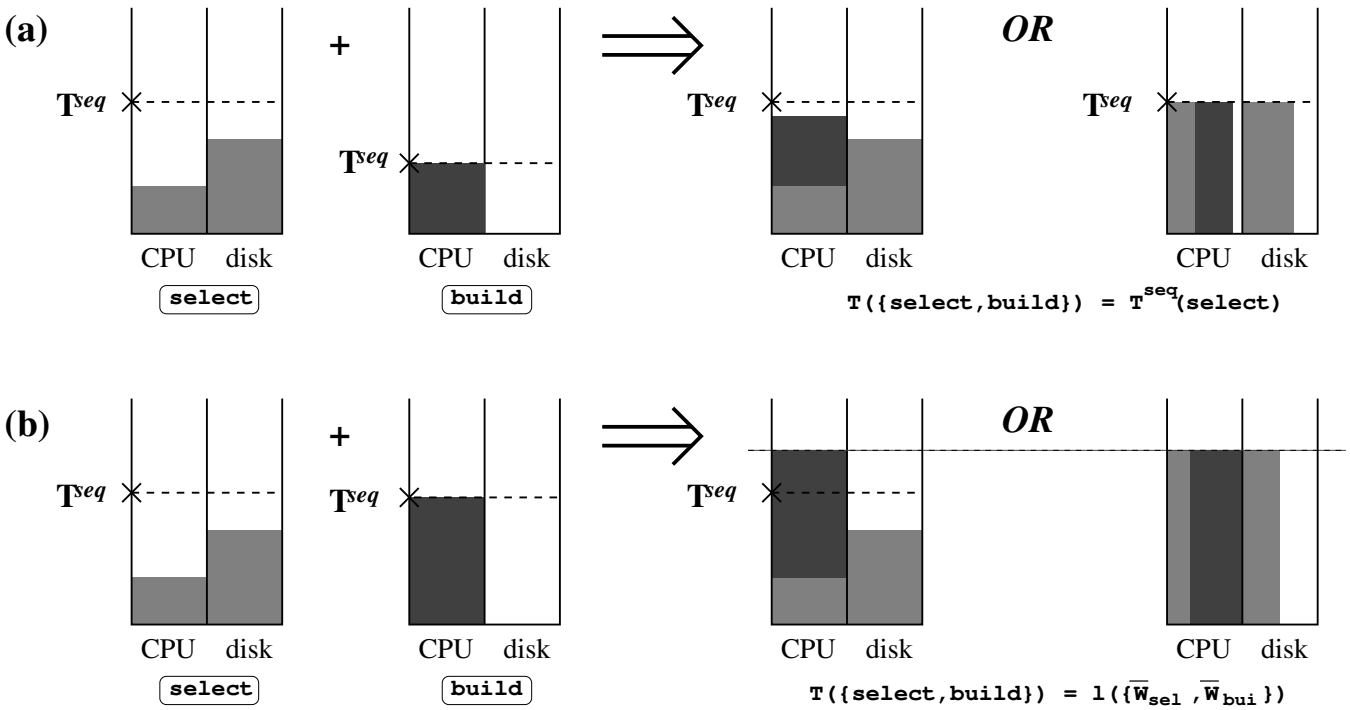

Figure 5: Packing the work vectors of compatible clones at a system site when (a) the slowest execution time can accommodate the total TS requirements; and, (b) a TS resource (CPU) becomes the bottleneck for the concurrent execution.

SCHED. Then, the total execution time for site B_j is simply the sum of the execution times for all compatible subsets in $S(B_j)$; that is,

$$T^{site}(B_j) \;=\; \sum_{S_i \in S(B_j)} T(S_i) \;=\; \sum_{S_i \in S(B_j)} \max\{ \max_{\overline{W} \in S_i^W} \{T^{seq}(\overline{W})\} \;,\; l(S_i^W) \}. \qquad (1)$$

Finally, the response time of the schedule SCHED is obviously determined by the longest running system site; that is, $T^{par}(\text{SCHED}, P) \;=\; \max_{1 \le j \le P}\{\; T^{site}(B_j) \}$.

4.2.2 Pipelined Parallelism

Pipelined parallelism introduces a *co-scheduling requirement* for query operators, requiring a collection of clones to execute in producer-consumer pairs using fine-grain/lock-step synchronization. The problems with load-balancing a pipelined execution have been identified in previous work [Gra93]. Compared to our model of a schedule for partitioned and independent parallelism (Definition 4.1), pipelined execution constrains the placement and execution of compatible clone subsets to ensure that all the clones in a pipe run concurrently – they all start and terminate at the same time [HM94]. As we suggested in the previous section, this means that it is no longer possible to schedule resources at one site independent of the others. Compatible subsets containing clones from the same (physical) operator pipeline *must* run concurrently. Furthermore, given that the scheduler does not modify the query plan, scheduling a pipeline is an "all-or-nothing" affair: either all clones will execute in parallel or none will. The implications of pipelined

parallelism for our scheduling problem will be studied further in Section 4.4 where a near-optimal solution will be developed.

4.3 Scheduling Independent Operators

In this section, we derive a lower bound on the optimal parallel execution time of independent operators (i.e., operators not in any pipeline) that accounts for both TS resource requirements and SS resource constraints. We then demonstrate that a heuristic based on Graham's LPT (Largest Processing Time) *list scheduling* method [Gra69] can guarantee near-optimal schedules for such operators.

Lemma 4.1 Let $\{\mathtt{op}_i, i = 1, \ldots, M\}$ be a collection of independent operators with respective degrees of partitioned parallelism $\{N_i, i = 1, \ldots, M\}$. Let S be the corresponding collection of clones and let $T^{max}(S)$ be the maximum stand-alone execution time among all clones in S; that is, $T^{max}(S) = \max_{i=1,\ldots,M}\{T^{max}(\mathtt{op}_i, N_i)\}$. If $T^{par}(\mathrm{OPT}, P)$ is the response time of the optimal execution on P sites then $T^{par}(\mathrm{OPT}, P) \geq LB(S, P)$, where

$$LB(S,P) \;=\; \max\{\, T^{max}(S)\,,\; \frac{l(S^W)}{P}\,,\; \frac{l(S^{TV})}{P}\,\}. \;\blacksquare$$

As with all other theoretical results in this paper, the proof of Lemma 4.1 is given in Appendix A.

The basic idea of our heuristic scheduling algorithm, termed OPSCHED, is to construct the partition of clones into compatible subsets (or, "shelves") incrementally, using a Next-Fit rule [CGJ84, CGJT80]. More formally, let n_i denote the number of compatible clone subsets that have already been mapped by OPSCHED to site B_i and let S_{i,n_i} denote the "topmost" compatible subset in B_i, i.e., the shelf containing the most recently scheduled clone at B_i. (Initially, $n_i = 0$ and $S_{i,0} = \phi$ for all i.) OPSCHED scans the list of independent clones to be scheduled in non-increasing order of execution time. At each step, the clone selected is placed in the site B_i with the minimum total execution time (i.e., bin "height") $T^{site}(B_i)$ (see Equation (1)). This placement is done as follows. If the clone can fit in the topmost compatible subset S_{i,n_i} without violating SS capacity constraints, then add the clone to S_{i,n_i} and update $T^{site}(B_i)$ accordingly. Otherwise, set $n_i = n_i + 1$, place the clone by itself in a new topmost subset S_{i,n_i}, and set $T^{site}(B_i)$ accordingly. The following theorem establishes an absolute performance bound of $d + 2s + 2$ for our heuristic.

Theorem 4.1 Given a set of *independent* operator clones S, OPSCHED runs in time $O(|S| \log |S|)$ and produces a schedule SCHED with response time

$$T^{par}(\mathrm{SCHED}, P) \;<\; d \cdot \frac{l(S^W)}{P} \;+\; 2s \cdot \frac{l(S^{TV})}{P} \;+\; 2 \cdot T^{max}(S) \;\leq\; (d + 2s + 2) \cdot LB(S, P). \;\blacksquare$$

Theorem 4.1 guarantees an *asymptotic performance bound*[6] of $d+2s$ for OPSCHED. This bound gives us a feeling for the performance of the algorithm when the optimal response time is much larger that the longest execution time of all clones and is a better measure of performance when $|S|$ is large [CGJT80, BS83].

[6] An asymptotic performance bound characterizes the behavior of an algorithm as the ratio of the optimal schedule response time to the longest job processing time goes to infinity, i.e., when $\frac{T^{par}(\mathrm{OPT},P)}{T^{max}(S)} \to \infty$.

Note that our scheduling algorithm combines the list scheduling method of Graham [Gra69] with the Next-Fit Decreasing Height (NFDH) shelf-based algorithm of Coffman et al. [CGJT80]. Essentially, the Next-Fit rule means that, at each step, only the topmost shelf is examined for placing a new clone. Using the more exhaustive First-Fit Decreasing Height (FFDH) shelving strategy in place of NFDH, we can combine our proof methodology with that of Coffman et al. to demonstrate a $d + 1.7s$ asymptotic performance bound for independent operator scheduling.

4.4 Scheduling with Pipelining Constraints

The co-scheduling requirement of pipelined operator execution introduces an extra level of complexity that OpSched cannot address, namely the problem of deciding whether a physical operator pipeline is *schedulable* on a given number of sites. Given a collection of clones in a pipeline, the schedulability question poses an $\mathcal{NP}$-hard decision problem that essentially corresponds to the decision problem of s-dimensional vector packing [CGJ84]. Thus, it is highly unlikely that efficient (i.e., polynomial time) necessary and sufficient conditions for pipeline schedulability exist. Note that no such problems were raised in the previous section, since the clones were assumed to be feasible (i.e., 1-granular) and executing independently of each other.

In this section, we demonstrate that λ-granularity (Definition 3.1) with $\lambda < 1$ for all operator parallelizations can provide an easily checkable sufficient condition for pipeline schedulability. (We use the term *λ-granular pipeline* to describe an operator pipeline in which all operator parallelizations are λ-granular.) Once schedulability is ensured, balancing the work and memory requirements of the pipeline across sites to minimize its response time still poses an $\mathcal{NP}$-hard optimization problem. We present a polynomial-time scheduling algorithm that is within a constant multiplicative factor of the response time lower bound for schedulable λ-granular pipelines. Further, we demonstrate that, using a level-based approach, our methodology can be extended to provide a provably near-optimal solution for multiple, independent physical operator pipelines.

4.4.1 Scheduling a Single λ-granular Pipeline

We present a near-optimal algorithm for scheduling a physical operator pipeline C consisting of λ-granular parallel operators, where $\lambda < 1$. Let S_C denote the collection of clones in C and, as always, let S_C^W and S_C^V be the corresponding sets of work and demand vectors, respectively. Also, let $T^{max}(S_C)$ denote the maximum stand-alone execution time among all clones in S_C. Note that, by our definitions, the pipeline C will require *at least* $l(S_C^V)$ sites for its execution (remember that each site has a capacity of 1 along all SS dimensions). The following lemma provides a sufficient condition for the schedulability of a λ-granular pipeline.

Lemma 4.2 The number of sites required to schedule a λ-granular pipeline C ($\lambda < 1$) is always less than or equal to $\frac{l(S_C^V) \cdot s}{1-\lambda}$. Furthermore, this bound is tight. ∎

Our heuristic, PipeSched, belongs to the family of list scheduling algorithms [Gra69]. PipeSched assumes that it is given a number of sites P_C that is sufficient for the scheduling of C, according to the condition of Lemma 4.2. The algorithm considers the clones in S_C in non-increasing order of their *work density ratio* $\frac{l(\overline{W}_i)}{l(\overline{V}_i)}$. At each step, the clone under consideration is placed in the site with the least amount of work that has sufficient SS resources to accommodate it; that is, clone $(\overline{W}_i, \overline{V}_i)$ is mapped to site B such

that $l(B^W)$ is minimal among all sites B_j such that $l(B_j^V \cup \{\overline{V}_i\}) \leq 1$. The full PipeSched algorithm is depicted in Figure 6.

Algorithm PipeSched(C, P_C)

Input: A set of λ-granular pipelined operator clones S_C and a set of P_C sites, where $P_C \geq \frac{l(S_C^V)\cdot s}{1-\lambda}$ (see Lemma 4.2).
Output: A mapping of the clones to sites that does not violate SS resource constraints. (Goal: Minimize response time.)

1. let $L =< (\overline{W}_1, \overline{V}_1), \ldots, (\overline{W}_N, \overline{V}_N) >$ be the list of all clones in *non-increasing* order of $\frac{l(\overline{W}_i)}{l(\overline{V}_i)}$.
2. for $k = 1$ to N do
   2.1. let $SB_k = \{B_j : l(B_j^V \cup \{\overline{V}_k\}) \leq 1\}$, i.e, the set of system sites with sufficient free SS resources to accommodate the k^{th} clone.
   2.2. let $B \in SB_k$ be a site such that $l(B^W) = \min_{B_j \in SB_k}\{l(B_j^W)\}$.
   2.3. place clone $(\overline{W}_k, \overline{V}_k)$ at site B and set $B^W = B \cup \{\overline{W}_k\}$, $B^V = B \cup \{\overline{V}_k\}$.

Figure 6: Algorithm PipeSched

Example 4.2: Consider a simple physical operator pipeline C comprising $N = 4$ clones with work (CPU+disk) and memory demands as follows: $(\overline{W}_1, \overline{V}_1) = ([10, 5], [0.2])$, $(\overline{W}_2, \overline{V}_2) = ([15, 0], [0.3])$, $(\overline{W}_3, \overline{V}_3) = ([7, 9], [0.3])$, and $(\overline{W}_4, \overline{V}_4) = ([2, 10], [0.35])$. These clones are obviously λ-granular with $\lambda = 0.35$. We now demonstrate how our PipeSched algorithm would schedule this pipeline on a set of $P_C = 2$ sites. First, observe that C is clearly schedulable on 2 sites, since $\frac{l(S_C^V)}{1-\lambda} = \frac{0.2+0.3+0.3+0.35}{1-0.35} = 1.77 \leq 2$. Also, note that the clone subscripts denote the ordering of the clones in non-increasing work density; that is, $\frac{l(\overline{W}_1)}{l(\overline{V}_1)} = \frac{10}{0.2} = 50 \geq \frac{l(\overline{W}_2)}{l(\overline{V}_2)} = 50 \geq \frac{l(\overline{W}_3)}{l(\overline{V}_3)} = 30 \geq \frac{l(\overline{W}_4)}{l(\overline{V}_4)} = 28.58$. (The tie between the first two clones is broken arbitrarily.)

Figure 7 depicts the series of steps taken by PipeSched and the final execution schedule produced. Note that, even for this very simple example, PipeSched is able to effectively balance the work and memory requirements of the clones across the two system sites. ∎

The following theorem establishes an asymptotic upper bound of $d(1 + \frac{s}{1-\lambda})$ on the worst-case performance ratio of our algorithm.

Theorem 4.2 Given a λ-granular pipeline C ($\lambda < 1$) PipeSched runs in time $O(|S_C| \log |S_C|)$ and produces a schedule SCHED with response time

$$T^{par}(\text{SCHED}, P_C) \;\leq\; d(1 + \frac{s}{1-\lambda}) \cdot \frac{l(S_C^W)}{P_C} + T^{max}(S_C) \;\leq\; [d(1 + \frac{s}{1-\lambda}) + 1] \cdot LB(S_C, P_C). \;\blacksquare$$

Note that the volume term of our lower bound (Lemma 4.1) does not come into the expression for the performance bound of PipeSched. This is because, by definition, all the clones in S_C must execute in parallel and thus $l(S_C^V) \leq P_C$. Thus, for a single pipeline, we always have $\frac{l(S_C^{TV})}{P_C} \leq T^{max}(S_C) \cdot \frac{l(S_C^V)}{P_C} \leq$

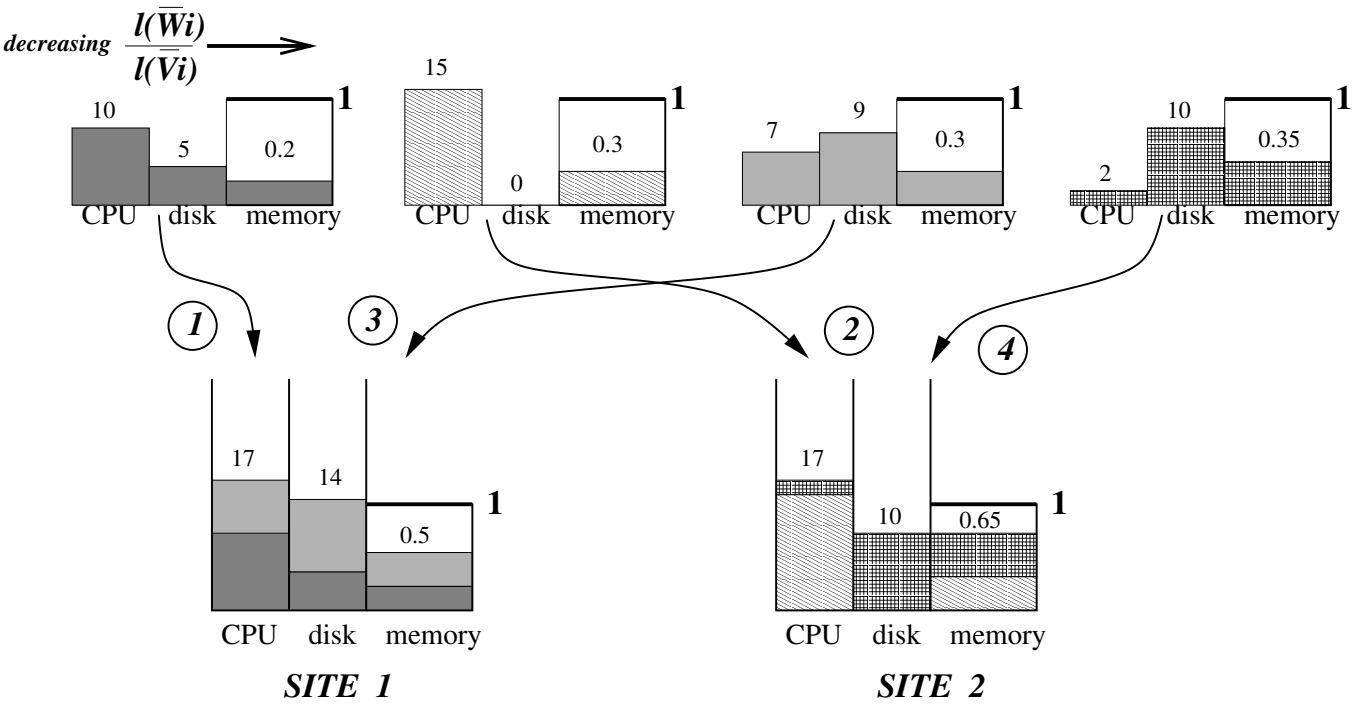

Figure 7: An example schedule produced by algorithm PipeSched.

$T^{max}(S_C)$; that is, the maximum clone execution time always dominates the average volume for a single physical operator pipeline.

The bound established in Theorem 4.2 clearly captures the granularity tradeoffs identified in Section 3. Increasing the degree of partitioned parallelism decreases both $T^{max}(S_C)$ and λ, thus allowing for a better asymptotic bound on the ratio and a smaller additive constant. On the other hand, it also increases the total amount of work $l(S_C^W)$ because of the overhead of parallelism. The importance of such work-space tradeoffs for parallel query processing and optimization has been stressed in earlier work as well [HFV96].

4.4.2 Scheduling Multiple Independent λ-granular Pipelines

The basic observation here is that our PipeSched algorithm presented in the previous section can be used to schedule any collection of independent pipelines as long as schedulability is guaranteed by Lemma 4.2.

Our algorithm for scheduling multiple independent pipelines uses a Next-Fit Decreasing Height (NFDH) policy [CGJT80] in conjunction with Lemma 4.2 to identify pipelines that can be scheduled to execute concurrently on P sites (i.e., in one layer/"shelf" of execution). PipeSched is then used for determining the execution schedule within each layer. The overall algorithm, termed LevelSched, is formally outlined in Figure 8.

The following theorem gives an upper bound on the worst-case performance ratio of LevelSched. Note that the co-scheduling requirement for the clones in a pipe implies that the total volume for all the clones in $\{C_1, \ldots, C_N\}$ is $l(S^{TV}) = l(\sum_1^N T^{max}(S_{C_i}) \cdot \sum_{\overline{v} \in S^V_{C_i}} \overline{v})$, since any clone in C_i will require its share of ss resources for at least $T^{max}(S_{C_i})$ time. The lower bound in Lemma 4.1 holds using the above definition of volume.

Algorithm LevelSched($\{C_1, \ldots, C_N\}$, P)
Input: A set of λ-granular operator pipelines $\{C_1, \ldots, C_N\}$ and a set of P sites.
Output: A mapping of clones to sites that does not violate SS resource constraints or pipelining dependencies. (Goal: Minimize response time.)

1. Sort the pipelines in non-increasing order of T^{max}, i.e., let $L =< C_1, \ldots, C_N >$, where $T^{max}(S_{C_1}) \geq \ldots \geq T^{max}(S_{C_N})$.
2. Partition the list L in k *maximal schedulable sublists:* $L_1 =< C_1, \ldots, C_{i_1} >$, $L_2 =< C_{i_1+1}, \ldots, C_{i_2} >$, ..., $L_k =< C_{i_{k-1}+1}, \ldots, C_N >$ based on Lemma 4.2. That is,

$$l(\cup_{C \in L_j} S_C^V) \leq \frac{P(1-\lambda)}{s} \quad \text{and} \quad l((\cup_{C \in L_j} S_C^V) \cup S_{C_{i_j+1}}^V) > \frac{P(1-\lambda)}{s}, \quad \text{for all } j = 1, \ldots, k-1.$$

3. for $j = 1, \ldots, k$ do
   3.1. call PipeSched($(\cup_{C \in L_j} C)$, P)

Figure 8: Algorithm LevelSched

Theorem 4.3 Given a collection of N independent λ-granular pipelines ($\lambda < 1$) comprising a set of clones S, LevelSched runs in time $O(N|S| \log P|S|)$ and produces a schedule SCHED with response time

$$T^{par}(\text{SCHED}, P) < d^2(1+\frac{s}{1-\lambda}) \cdot \frac{l(S^W)}{P} + \frac{2s^2}{1-\lambda} \cdot \frac{l(S^{TV})}{P} + T^{max}(S) \leq [d^2(1+\frac{s}{1-\lambda}) + \frac{2s^2}{1-\lambda} + 1] \cdot LB(S, P). \blacksquare$$

It is important to note that the lower bound estimated in Lemma 4.1 will, in most cases, significantly underestimate the optimal response time since it assumes that 100% utilization of system resources is always possible *independent of the given task list*. Thus, the quadratic multiplicative constants in Theorem 4.3 reflect only a worst case that is rather far from the average. This fact has been verified through extensive experimentation using cost model computations in our earlier work [GI97].

4.5 Handling Blocking Constraints and On-Line Task Arrivals

Our discussion so far has focused on the case of scheduling a collection of one or more *independent* physical operator pipelines. In this section, we propose heuristic extensions that allow our scheduling algorithms to handle *blocking constraints* (and the corresponding execution dependencies (Section 2.1)) that arise in *logical* operator (e.g., hash-join) pipelines and bushy query execution plans. Our proposed extensions deal with the two key aspects of query operator scheduling, namely (1) determining the degree of partitioned parallelism (i.e., number of clones) for interdependent operators, and (2) mapping operator clones to sites so that execution dependencies are respected.

• **Determining the Degree of Partitioned Parallelism.** Blocking parent-child operator pairs with a disk- or memory-materialization dependence are always constrained to run on the exact same set of system sites. This is the case, for example, with a `store` operator at the end of a query task and the `scan` (or `select`) operator in its parent query task (disk-materialization dependence) or a `build` operator

and its parent `probe` (memory-materialization dependence). In such scenarios, determining the degree of parallelism for the child operator independently of its parent may result in poor parallelization decisions. As an example, the amount of work for a `build` depends on the inner/build relation size whereas the amount of work for a `probe` depends primarily on the size of the outer/probe relation. Fixing the degree of parallelism for a hash-join based on just one of these two operations is probably a bad idea, since the inner and outer relation sizes can be vastly different.

Given such a pair of execution-dependent operators $\texttt{op}_1$ (parent) and $\texttt{op}_2$ (child), we propose setting their (common) degree of partitioned parallelism based on an extension of the definition for λ-granular CG_f executions (Definition 3.3). The basic idea is to use the total processing area $\mathcal{A}_p(\texttt{op}_1) + \mathcal{A}_p(\texttt{op}_2)$ and total communication area $\mathcal{A}_c(\texttt{op}_1, N) + \mathcal{A}_c(\texttt{op}_2, N)$ of the operator pair in defining the communication granularity of the execution, while ensuring that the memory granularity constraint λ is enforced for the (floating) child operator $\texttt{op}_2$. More formally, given the system-wide granularity parameters f and λ, we set the degree of parallelism of $\texttt{op}_1$ and $\texttt{op}_2$ to the maximal N such that $(\mathcal{A}_c(\texttt{op}_1, N) + \mathcal{A}_c(\texttt{op}_2, N)) \leq f' \cdot (\mathcal{A}_p(\texttt{op}_1) + \mathcal{A}_p(\texttt{op}_2))$, where f' is the minimum value larger than or equal to f such that $\mathcal{V}(\texttt{op}_2, N) \leq \lambda$. Combining the work demands allows more effective parallelization decisions to be made for execution-dependent operator pairs.

• **Mapping Clones to Sites.** Scheduling operator clones from arbitrary bushy query execution plans must ensure that all execution dependencies specified by the blocking edges in the corresponding query task tree (Figure 2(c)) are satisfied. Our LevelSched algorithm can be readily extended to handle *disk-materialization dependencies* by simply ensuring that the (sorted) ready list of tasks L always contains the collection of query tasks that are ready for execution, i.e., they are not blocked waiting for the materialization of some intermediate result on disk from some other (descendant) task in the task tree. For *memory-materialization dependencies*, care must be taken to ensure that when memory-dependent collections of query tasks are present in adjacent levels of the task tree (e.g., a pipeline of `probe`s and all its child `build` pipelines) then (1) the clones of all child tasks are executed in parallel, and (2) the clones of all parents are executed *immediately afterwards* using the exact same sites and memory resources. Given this very tight coupling of memory-dependent physical operators, we propose treating such parent-child pairs as a single, "combined" operator with a work vector that captures the total TS requirements for both phases. (Of course, the SS demand vector remains the same throughout the execution of the combined operator.) This heuristic method tries to account for the tight execution dependencies between memory-dependent operator pairs and achieve better overall load-balancing by essentially "merging" the two corresponding execution shelves into one. As an example, in the query task of Figure 2(c), our proposed method merges the `build` pipelines T3 and T4 into their parent `probe`s in T5. All this is achieved by modifying LevelSched to produce a new algorithm TreeSched as follows (see Figure 8):

1. Any sibling physical operator pipelines with a *memory-materialization dependence* to their parent in the task tree are merged into their respective parent operators in the parent task (as described above). This essentially gives rise to a "coarser" parent pipeline that is treated as a unit, i.e., the way individual pipelines are treated in LevelSched. For the purposes of this algorithm, assume that the term 'pipeline' is interpreted as such a unit.

2. Initially, the input set of pipelines $\{C_1, C_2, \ldots\}$ contains exactly the tasks at the leaf nodes of the query task tree.

3. After Step 3.1, determine the set of tasks $\mathbf{C}$ that have been enabled (i.e., are no longer blocked) because of the last invocation of PipeSched. If $\mathbf{C} \neq \emptyset$, then merge the tasks in $\mathbf{C}$ into the ready list L and go to Step 2. Otherwise, continue with the next invocation of PipeSched.

The exact same idea of dynamically updating and partitioning the ready list L can be used to handle *on-line* task arrivals in a dynamic or multi-query environment. Basically, newly arriving query tasks are immediately merged into L to participate in the partitioning of L into schedulable sublists right after the completion of the current execution layer. Thus, our layer-based approach provides a uniform scheduling framework for handling intra-query as well as inter-query parallelism.

Deriving theoretical performance bounds for our scheduling algorithms in the presence of blocking constraints and execution dependencies is a very difficult problem that continues to elude our efforts. The difficulty stems from the interdependencies between different execution layers: scheduling decisions made at earlier layers can impose data placement and execution constraints on the layers that follow. We leave this problem open for future research.

5 Performance Evaluation Using Detailed Simulation

In our earlier work [GI97], we have presented the results of several experiments conducted using cost model computations for various query operators in order to examine the *average-case performance* of our algorithms compared to the optimal solution. These results have demonstrated that the average-case performance of our multi-resource scheduling algorithms is much closer to optimal than what would be predicted by the pessimistic worst-case bound of Theorem 4.3. Thus, on the average, our algorithms produce very effective, near-optimal packings of memory and work requirements over the system sites.

In this paper, we present a performance study of our parallel query scheduling algorithms conducted using a detailed simulation model for a shared-nothing parallel database system [Bro94]. The simulator is written in the CSIM/C++ process-oriented simulation language [Sch90b] and is based on the architecture of the Gamma parallel database machine [DGS+90].

Figure 9(a) gives a high-level overview of our experimentation procedure. Briefly, the scheduling algorithms investigated in our simulation study receive as input a set of query execution plans along with schema and system configuration information (e.g., declustering of base relations, sizes of intermediate results, disk and CPU characteristics) and produce as output an execution schedule (i.e., a mapping of query plan operators onto system sites). The scheduling algorithm makes its mapping decisions off-line, as would be the case if it were part of a complete query optimizer, based on its input and (possibly) on cost model computations. For example, our multi-resource scheduler makes its scheduling choices using a cost model to determine the work and demand vectors for operator clones and taking the componentwise sum of vectors mapped to a specific site to estimate the expected TS and SS resource loads at that site. The effectiveness of the execution schedule produced is determined by "feeding" the schedule to our simulator which *actually executes* it on top of our detailed model of a shared-nothing parallel database system and outputs the response time of the execution.

The following section describes the various components of our simulation model in detail. We then go on to describe our experimental testbed and discuss the simulation results.

5.1 Execution Environment

5.1.1 Simulator Query Processing Architecture

Our simulator models a typical shared-nothing system that consists of a collection of *query processing sites*, each comprising one or more CPUs, memory, and one or more disk drives. The sites use an interconnection network for all communication. Query plans are submitted for execution from external terminals to a

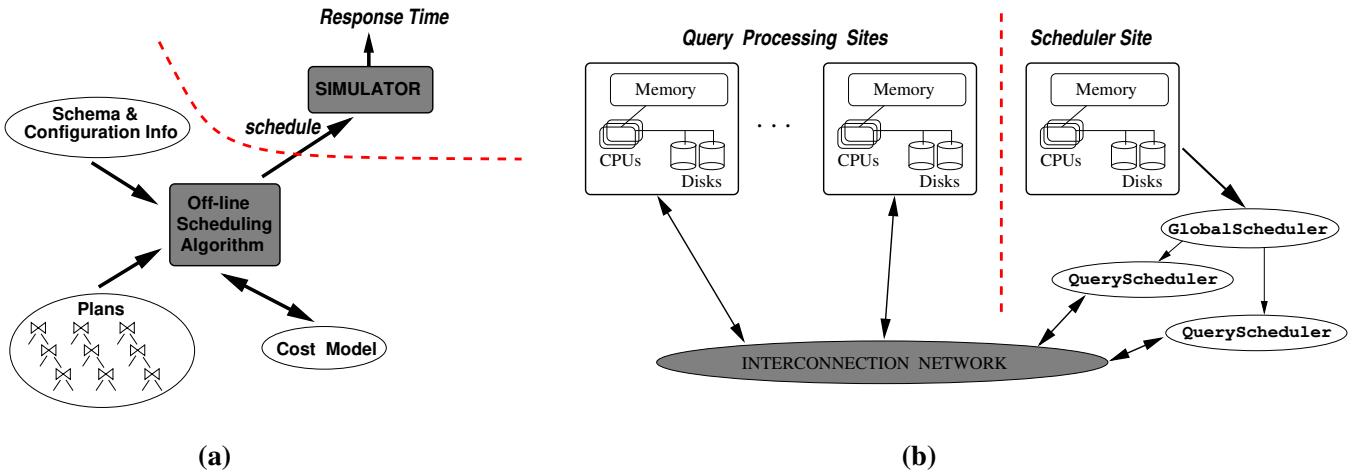

Figure 9: (a) Experimentation procedure. (b) Simulator query processing architecture.

`GlobalScheduler` process running on a dedicated *scheduler site* that lies outside the set of query processing sites in the system. The `GlobalScheduler` determines a mapping of plan operators to sites, allocates the appropriate resources for the plan, and finally initiates a `QueryScheduler` process on the scheduler site that is responsible for executing the plan to completion. For each operator to be executed, the `QueryScheduler` creates an execution thread on every site executing a clone of the operator. For each such thread, appropriate communication channels are set up for the exchange of control messages between the `QueryScheduler` and the thread (e.g., transition from the `build`-phase to the `probe`-phase of a join pipeline, split table broadcasting, and "clone done" messages) and data messages between threads of neighboring operators in the plan. When a plan is run to completion the corresponding `QueryScheduler` dies and a reply message is sent to the submitting terminal. The `GlobalScheduler` may then choose to activate another query plan that had been waiting for some system resources to be released. Figure 9(b) provides a high-level view of the query processing architecture in our simulation environment. It is the same one assumed for our analytical results (Figure 1) with respect to where plans are to run (to the left of the vertical dashed line), complemented with one more site where the scheduler runs to produce the appropriate schedule (to the right of the vertical dashed line).

5.1.2 Hardware and Operating System Characteristics

Query processing sites comprise one or more CPUs, a buffer pool of 8 KByte pages, and one or more disk drives. CPU cycles are allotted using a round-robin policy with a 5 msec timeslice. The buffer pool models a set of main memory page frames whose replacement is controlled by a 3-level LRU policy extended with *"love/hate"* hints [HCL+90]. In addition, a memory reservation mechanism under the control of the `GlobalScheduler` process allows memory to be reserved for a particular query operator. This mechanism is employed to ensure that hash table page frames allocated to join operators will not be stolen by other operators.

The simulated disks model a Seagate Cheetah 36 Model ST136403FC (36.4 GByte, 3.5") [Sea], a "state-of-the-art" SCSI disk drive. This disk drive provides a 1024 Kbyte cache that is divided into sixteen 64 KByte cache contexts to effectively support up to sixteen concurrent sequential prefetches. In the disk

model, which slightly simplifies the actual operation of the disk, the cache is managed as follows. Each I/O request, along with the required page number, specifies whether or not prefetching is desired. If prefetching is requested, eight additional pages are read from the disk into a cache context as part of transferring the originally requested page from disk to memory. Subsequent requests for one of the prefetched pages can then be satisfied without incurring an I/O operation. A simple round-robin policy is used to allocate cache contexts if the number of concurrent prefetch requests exceeds the number of available cache contexts. The disk queue is managed using an elevator algorithm [SG98].

Our simulator models the interconnection network as a set of point-to-point connections between every pair of system sites. The speed of each point-to-point link is set to 200 Mbits/sec, which ensures that network speed is not the bottleneck in our experiments. (This is in agreement with most real parallel systems (e.g., iPSC/2, Hypercube, Paragon), where the interconnect rarely is the bottleneck resource during parallel processing.) However, the details of the particular interconnect technology are not modeled. The cost of message-passing is captured as follows: our simulator charges for the time-on-the-wire (depending on network latency, message size, and network bandwidth) as well as CPU instructions for networking protocol operations, which consist of (1) a fixed cost per message, and (2) a per-byte cost based on the size of the message. The CPU costs for messages are paid both at the sender and at the receiver. Table 2 summarizes the execution environment parameters employed in our simulations. The CPU instruction counts for various database operations are based on earlier simulation studies [Meh94, PI96].

Configuration/Site Parameters	**Value**
Number of Sites	16 – 96
CPUs per Site	1 × 100 MIPS
Disks per Site	2 × Cheetah-36
Memory per Site	64, 96, 128 MB
Page Size	8 KB
Network Bandwidth	200 Mbits/sec
Minimum Wire Delay	0.1 msec
Disk Seek Factor	0.1414
Disk Rotation Time	5.98 msec
Disk Settle Time	0.5 msec
Disk Transfer Rate	29 MB/sec
Disk Cache Context Size	8 pages
Disk Cache Size	16 contexts
Disk Cylinder Size	560 pages
Buffer Manager	3-level LRU

CPU Cost Parameters	**No. Instr.**
Initiate Select	20000
Initiate Join	40000
Initiate Store	10000
Terminate Select	5000
Terminate Join	10000
Terminate Store	5000
Read Tuple	300
Write Tuple into Output Buffer	100
Probe Hash Table	200
Insert Tuple in Hash Table	100
Hash Tuple Using Split Table	500
Apply a Predicate	100
Copy 8K Message to Memory	10000
Message Protocol Costs	1000

Table 2: Simulation Parameter Settings

5.2 Experimental Testbed and Methodology

Scheduling Algorithms. Our experiments have focused on three different scheduling algorithms:

- TreeSched: Our level-based, multi-resource scheduling algorithm for query task trees, observing blocking constraints and execution dependencies across execution layers (Section 4.5).
- Hier: A one-dimensional, processor allocation algorithm based on the hierarchical method developed by Wolf et al. [WTCY94, WTCY95] for scheduling multiple parallel queries on a set of processors.

(Since the original method did not account for pipelining dependencies or memory constraints, we had to adapt it appropriately.) Briefly, HIER "collapses" each right-deep join pipeline into a single parallel job and incrementally explores different possible allotments of processors to individual joins starting from a "minimum work" allotment. At each step, a list schedule based on the current processor allotments is constructed and is used to identify the pipeline with the maximum "processor waste". Then, the degree of parallelism for the maximum-waste job is increased by one by giving one more processor to its most time-consuming join. This process is repeated until a list schedule with zero waste is produced, and the best (i.e., shortest makespan) schedule found so far is chosen.

- ZETA: The "standard", simple query scheduling algorithm that has been built-in the `GlobalScheduler` process of our simulator. ZETA simply runs each join operator at the set of sites where its left input (i.e., `build`-relation) is declustered.

Note that the search process for the HIER algorithm is always guaranteed to terminate, since it will eventually reach an allotment where every pipeline is allotted all sites in the system, in which case the "processor waste" is zero for all jobs. This, however, also means that, in the worst case, the HIER scheduler has to build $O(N \cdot P)$ list schedules, where N is the number of join pipelines and P is the number of sites. Thus, the running time complexity of HIER can be significantly higher than that of our TREESCHED algorithm.

Data Placement Strategies. For all three scheduling algorithms, we experimented with various possible data placement strategies for the base relations of our workload queries:

- **Declust**: Every base relation is horizontally partitioned across all sites in the system.
- **Declust-1/4**: Every base relation is horizontally partitioned across the same subset of the system sites, comprising 1/4th of the entire system.
- **NoDeclust**: A base relation resides on a *single* system site, chosen randomly for every relation from the set of all sites.
- **NoDeclust-1/4**: A base relation resides on a *single* system site, chosen randomly for every relation from the same subset of the system sites, comprising 1/4th of the entire system.
- **Random**: Both the degree of declustering and the actual sites storing a relation are selected randomly from the underlying set of sites. To avoid partitioning small relations over too many sites (which is typically avoided in real systems), we placed an upper bound on the degree of declustering of a relation that is proportional to its size. (The upper bound for the largest relation in the system is the total number of sites.) The actual degree of declustering was then randomly chosen between 1 and that upper bound.
- **QueryBasedDeclust**: The degree of declustering for a base relation is equal to the minimum number of sites required to hold the entire `build` hash table in main memory. Furthermore, the actual set of sites given to a `build` relation is carefully selected so that *it does not overlap* with any of the homes of the other `build` relations in the same query.

Obviously, **Declust** and **NoDeclust** represent the two extreme choices in the space of data declustering strategies. The **1/4**-versions of these strategies were chosen to represent situations in which input queries will experience hotspots due to data placement choices. Although the choice of the fraction of the system

used for mapping data (1/4) was arbitrary, we believe that our experimental observations will remain valid for other system fractions. The **Random** placement strategy tries to relieve query hotspots by mapping relation fragments to randomly selected sites. Such hotspots may still occur, however, based on the actual relations accessed by a query. For example, it may very well happen that the homes of the four `build` relations in a 4-join right-deep query overlap at one or more system sites. These sites will then be the bottlenecks for the `build` phase of the pipe. It is exactly such bottlenecks that our final and more sophisticated data placement strategy, **QueryBasedDeclust**, tries to avoid by explicitly examining the set of queries to be executed. Of course, determining the exact query workload may not always be possible (e.g., in an ad-hoc querying environment).

Query Workloads. We executed all three scheduling algorithms over the same set of input right-deep hash-join query plans (submitted from the external terminals at time 0), for each of the aforementioned base relation placement strategies. Although our analytical results are completely general, we focused our simulations on hash-joins only, since it is well-known that they offer several advantages over other join algorithms in a parallel execution environment, particularly because of the opportunities they present for pipelining [CLYY92, DGS+90, HCY94, LCRY93]. For each workload, we generated right-deep query plans of sizes 40, 60, and 80 over randomly selected base relations. Given the total number of joins in the workload, different right-deep plan size combinations were tested (e.g., our 40-join workloads included one workload with five 4-join pipelines and ten 2-join pipelines, one workload with ten 4-join plans, etc.). For each workload combination, we executed five randomly generated workloads (by randomly selecting the participating base relations) and recorded the observed simulation times for each of the five runs. The results presented in the next section are indicative of the results obtained for all randomly generated workloads and different combinations of pipeline sizes. For the experiments presented in this section, there was *no base relation shared across query plans*, i.e., each relation was part of at most one join in a workload. The purpose of this was to mitigate the effects of the buffer management algorithm on our scheduling results. We plan to explore these effects in future work.

Our simulator implements a hybrid hash-join processing algorithm for right-deep trees, so it is possible to schedule each join in a pipe to be executed with as little as the square root of its maximum memory demand [Sch90a, Sha86]. Of course, using less memory implies that the join pipeline is executed in multiple passes. This is never the case for LevelSched, since demand vectors are always determined using the maximum (i.e., one-pass) memory demand for a join. Since our initial results showed that multi-pass join pipelining can really hurt the overall response time performance for a workload, we also tried to avoid multi-pass pipeline executions for Hier by starting with a "minimum work" allotment that gives each join a sufficient number of processors for keeping its entire `build` hash table in memory. Note, however, that this may not be possible under certain ("small") configurations, since Hier does *not* allow joins in a pipeline to share processors. In such scenarios, we allocated sites to joins in the pipeline in proportion to their total memory requirement. For Zeta, we implemented a scheduling switch that forces the `GlobalScheduler` process not to start the execution of a query until its maximum memory requirement can be satisfied. Once again, however, this may be impossible given the degree of declustering of the `build` relations.

All our right-deep plans included a final step of storing the result relation across all sites using a single disk at each site. We decided to use this default strategy for `store`'s rather than treating them as floating operators since for many real-world queries the results do not actually need to be stored on disk, so we wanted to minimize the impact of `store`'s on the relative quality of the execution schedules. We also employed the simulator's built-in assumption of uniform join attribute distributions and decided to leave the study the effectiveness of the schedules generated by the various algorithms in the presence

of data skew to future work. However, we firmly believe that skew can only increase the gains of our scheduling algorithm over both HIER and ZETA. The reason is that, unlike TREESCHED, (a) HIER's list scheduling scheme is based on the assumption of uniform partitioning of an operator's work among its clones [WTCY94, WTCY95], and (b) by its definition, ZETAis not allowed to re-balance the tuples of a `build` relation across sites.

Our basic comparison metric was the observed *response time* for a given query workload. The values of our database schema and workload parameters are summarized in Table 3. Both the coarse granularity parameter f and the memory granularity parameter λ were varied over a wide range of possible values in an effort to determine their impact on the performance of LEVELSCHED. (The range of values for the f parameter was decided based on the experimental observation of some actual values for operator processing and parallelization costs during the operation of our simulator.) Although "fine tuning" f and λ is a non-trivial problem, we will discuss some of the insights obtained from our experimental findings in the next section. The *join probability* is the factor multipled by the cardinality of the right input to determine the result cardinality of a join, and the *projection factor* indicates the percentage of the sum of the left and right input tuple widths that should be retained in the result. We kept these join parameters fixed at 1.0 and 0.5, respectively during all our experiments.

Schema/Workload Parameters	**Value**
Tuple Size	200 bytes
Base Relation Size	$10^4 - 10^6$ tuples
Join Attribute Distribution	Uniform
Workload Size (Total No. of Joins)	40, 60, 80 joins
Right-Deep Query Size	2, 4, 8 joins
Coarse Granularity Parameter (f)	$0.001 - 0.1$
Memory Granularity Parameter(λ)	$0.1 - 0.75$
Join Probability	1.0
Projection Factor	0.5

Table 3: Database Schema and Workload Parameters

5.3 Experimental Results

The objective of our empirical study was twofold: (1) to examine the impact of the clone granularity parameter values (f and λ) on the performance of our TREESCHED algorithm, and (2) to demonstrate the benefits of our multi-resource scheduling strategies compared to earlier one-dimensional or "naive" approaches. The major findings of our study can be summarized as follows.

- **Insensitivity of TREESCHED to Clone Granularity Parameters.** The problem of "optimally" tuning the clone granularity parameters (f and λ) is obviously very complex and depends on a multitude of factors. However, our results demonstrate that the performance of the schedules produced by our TREESCHED algorithm is relatively insensitive to the exact choice of values for the granularity parameters, as long as certain extreme or pathological scenarios for operator parallelization are avoided.

- **Improved Schedule Quality Compared to HIER and ZETA.** The quality of the schedules produced by our TREESCHED algorithm is, in general, superior to that obtained by either the "naive"

ZETA scheduler or the one-dimensional HIER algorithm. This demonstrates TREESCHED's ability to effectively balance the query workload across the system sites and verifies the benefits of TREESCHED's multi-dimensional approach for balancing the utilization of different resources within the same site. As our results indicate, the need for effective multi-resource scheduling is more evident in smaller, resource-limited system configurations.

We believe that our simulation results establish the practical importance of multi-resource scheduling for effective compile-time parallelization of query execution plans. Further, the fact that the quality of the schedules produced by our algorithm is relatively insensitive to the exact values of its input granularity parameters implies that our techniques should be easy to adapt and tune for different real-life parallel query optimizers.

5.3.1 Effect of Clone Granularity Parameters on the Performance of TREESCHED

Selecting "optimal" values for the coarse granularity parameter f and the memory granularity parameter λ of the TREESCHED algorithm is a obviously a very difficult problem. This is because, the optimal choice depends on a number of parameters and characteristics of the underlying hardware and query processing architecture, such as the CPU overhead of messages and clone synchronization, clone startup costs, and the speed of the interconnection network. Further, even for a fixed set of architectural parameters the impact of the granularity parameters on the observed workload response time can depend on characteristics of the workload itself, such as the size of the workload or the initial placement of base relation data. As an example, higher parallelization overheads (e.g., larger values for f) are obviously less desirable as the level of multiprogramming (i.e., number of distinct queries) in the system increases. Finally, as explained in Section 3.2, it is important to strike a balance between the overhead of parallelism and the ability to effectively "pack" memory demands at system sites. The implication is that it is probably a very hard problem to discover settings for f and λ that are "universally optimal". As our experimental results demonstrate, however, for our simulator configuration, the performance of our TREESCHED algorithm is stable and relatively insensitive to the exact choice of values for the granularity parameters. Based on our experimentation with other system configurations, we believe that this is a more general phenomenon. That is, we expect the performance of our multi-resource scheduling algorithm to be reasonably stable over a range of choices for the f and λ parameters, as long as certain pathological scenarios for operator parallelization are avoided. We try to provide some general guidelines for avoiding such scenarios at the end of this section.

Figure 10 depicts the response times of the execution schedules obtained by TREESCHED under four of the six data placement policies considered in our simulations. (The results for **Declust-1/4** and **NoDeclust-1/4** are very similar to those for **Declust** and **NoDeclust**, respectively.) Results are shown for a 60-join workload consisting of fifteen 4-join right-deep queries, for various values of the coarse granularity parameter f, with 64 MBytes of main memory per site, and a memory granularity parameter $\lambda = 0.2$. Nearly identical results were also obtained with different values of λ and different memory sizes. Our results clearly show that the performance of the schedules produced by our TREESCHED algorithm remains essentially unaffected by the specific choice of the coarse granularity parameter f. (The differences observed are typically in the range of 1-3%.) There are two main reasons for this. First, in our implementation of TREESCHED, we use the earlier results of Wilschut et al. [WFA92] to estimate and enforce an upper bound on the degree of effective partitioning for each floating operator. Thus, even for very large values of f, our TREESCHED implementation makes sure that the assigned degree of operator parallelism

never exceeds the "optimal number of clones" for the operator, as estimated by the formula of Wilschut et al. [WFA92]. Second, and perhaps most importantly, the startup and synchronization overhead of parallelism turns out to be relatively small, and thus higher numbers of join clones do not imply performance penalties for the schedule. This is mainly due to the high speed (100 MIPS) of site CPUs in our simulated system, which basically implies that the CPU overhead of setting up and synchronizing a few tens of clones is insignificant compared to the amount of query processing work. (This effect is also evident from the good performance of ZETA schedules under the fully-declustered **Declust** data placement policy even for "large" simulator configurations (Section 5.3.2).) Note that, in experiments for configurations with lower CPU speeds (10 MIPS) we found that such overheads can become significant when the number of sites exceeds 32. For such scenarios, more careful tuning of the coarse granularity parameter f may be required.

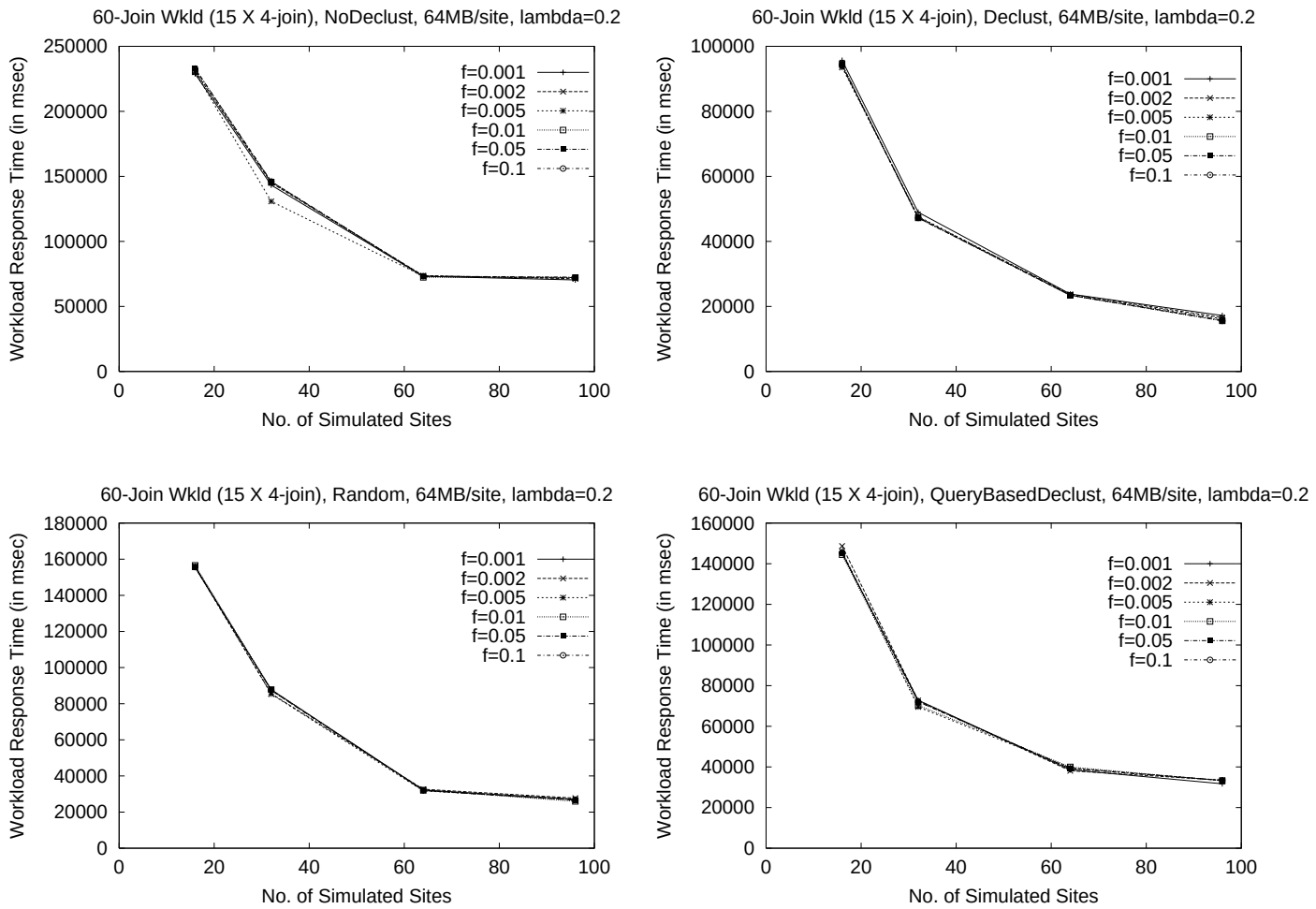

Figure 10: Effect of the coarse granularity parameter f on the performance of TREESCHED for a (15 X 4-join) workload under **NoDeclust**, **Declust**, **Random**, and **QueryBasedDeclust** data placement with 64 MBytes per site. ($\lambda = 0.2$)

The effect of different values for the memory granularity parameter λ on the performance of LEVELSCHED for the same 60-join workload is depicted in Figure 11 for **NoDeclust**, **Declust**, **Random**, and

QueryBasedDeclust data placement, a coarse granularity parameter of $f = 0.005$, and 64 MBytes of main memory per site. The results for **Declust-1/4** and **NoDeclust-1/4** are almost identical to those for **Declust** and **NoDeclust**, respectively, while very similar results were also obtained for different values of f and different memory sizes. The main observation again is that the performance of our TreeSched algorithm basically remains stable under different settings of the memory granularity parameter λ. (The observed differences are, once again, typically in the 1-3% range.)

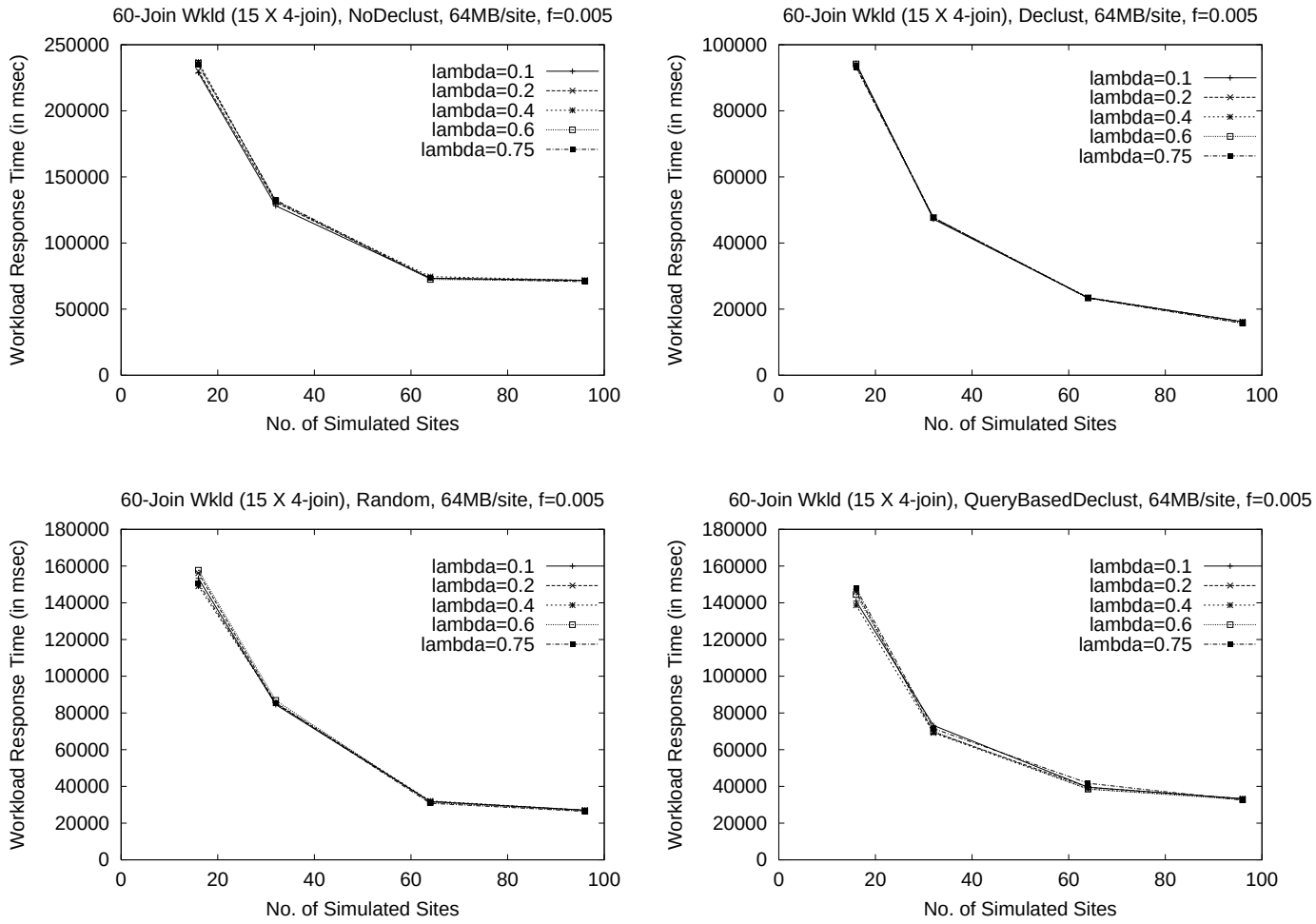

Figure 11: Effect of the memory granularity parameter λ on the performance of TreeSched for a (15 X 4-join) workload under **NoDeclust**, **Declust**, **Random**, and **QueryBasedDeclust** data placement with 64 MBytes per site. ($f = 0.005$)

General Guidelines for Clone Granularity. Our simulation results show that the performance of our TreeSched algorithm is stable and relatively insensitive to the exact choice of granularity parameters, for a reasonably wide range of possible values. There are, however, some general guidelines that should be adhered to when selecting f and λ values in order to avoid certain pathological scenarios with respect to operator parallelization. Abstractly, f tries to put an upper bound on the relative overhead of parallelization (compared to an operator's work), whereas λ determines how "thin" the memory demands of

clones are going to be. Based on our definition of λ-granular CG_f executions, we identify two "dangerous" scenarios that should probably be avoided when trying to choose reasonable (ranges of) values for f and λ. (Of course, such "reasonable" choices depend on system parameters and can be determined with some preliminary experimentation.)

- **Very large f or very small λ** typically imply that operators are "overpartitioned" to very thin clones, resulting in high parallelization overheads and possible slowdowns in execution time. This effect was not observed for our default simulator configuration setting (due to the reasons outlined above), but it did show up for certain configurations with lower CPU speeds.
- **Very large λ and very small f** typically imply that operators are assigned very low degrees of parallelism which may be too conservative to exploit the available system resources. An example of such an extreme case is depicted in Figure 12 for two distinct data placement strategies (**Declust** and **Declust-1/4**) and our default simulator settings with $f = 0.00001$ and $\lambda = 0.75$. As expected, under this fairly restrictive combination of values for f and λ, TreeSched is unable to appropriately leverage the additional resources as the system size grows.

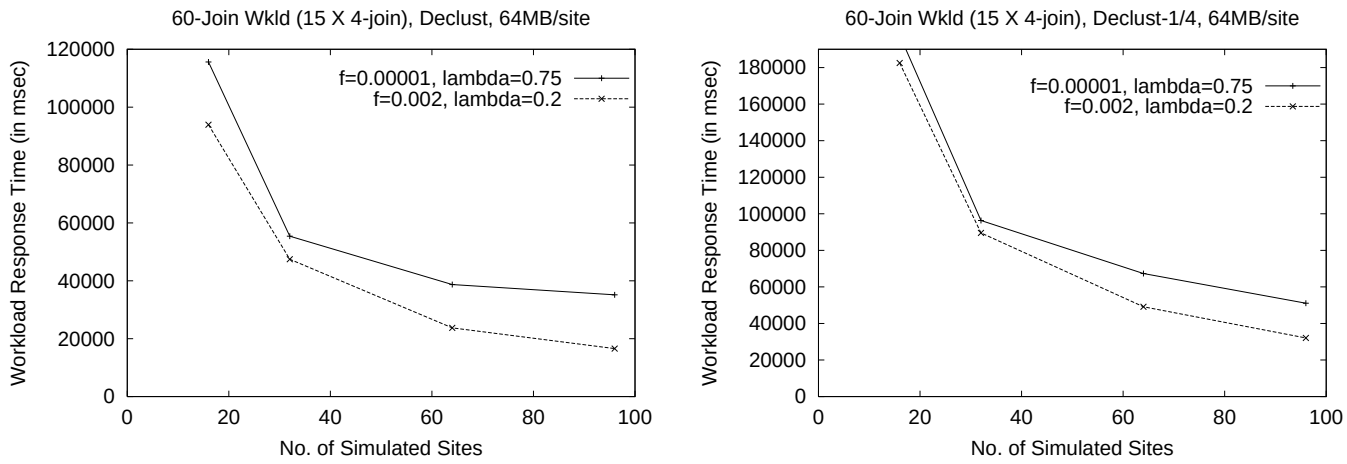

Figure 12: Effect of overly conservative combination of granularity parameters ($f = 0.00001$ and $\lambda = 0.75$) on the performance of TreeSched for (a) **Declust** and (b) **Declust-1/4**.

5.3.2 Comparison of Scheduler Performance

Figures 13–16 depict the performance of the three scheduling algorithms (TreeSched, Hier, and Zeta) for the various data placement strategies considered in our simulation study and site memory sizes of 64 and 96 MBytes. These results were obtained for a 60-join workload consisting of five 8-join queries and ten 2-join queries, assuming the default granularity parameter values of $f = 0.002$ and $\lambda = 0.2$.

Figure 13 gives the response times of the schedules produced by the TreeSched and Hier algorithms for the **NoDeclust** and **NoDeclust-1/4** data placement policies. As expected, both algorithms were clear winners over the naive Zeta strategy for our two "No-Declustering" policies — the response times for the Zeta schedules ranged from 4 to 16 times the corresponding response times for our TreeSched

scheduler, and are therefore omitted from the plots in Figure 13. The main reason for this large gap in performance is that, unlike ZETA, both TREESCHED and HIER manage to balance the memory demand for `build` relations across the system without being restricted by the specific sites at which base relations reside on disk. Between the two "smarter" schedulers, TREESCHED exploits its multi-dimensional model of operator costs to better balance the utilization of CPU and disk resources across system sites leading to improvements of up to 35-40% for certain configurations over the one-dimensional HIER schedules. Note that the performance differences between the two schedulers are typically diminished as the size of the system is increased. Obviously, effective resource scheduling becomes less critical once there is an abundance of resources in the system.

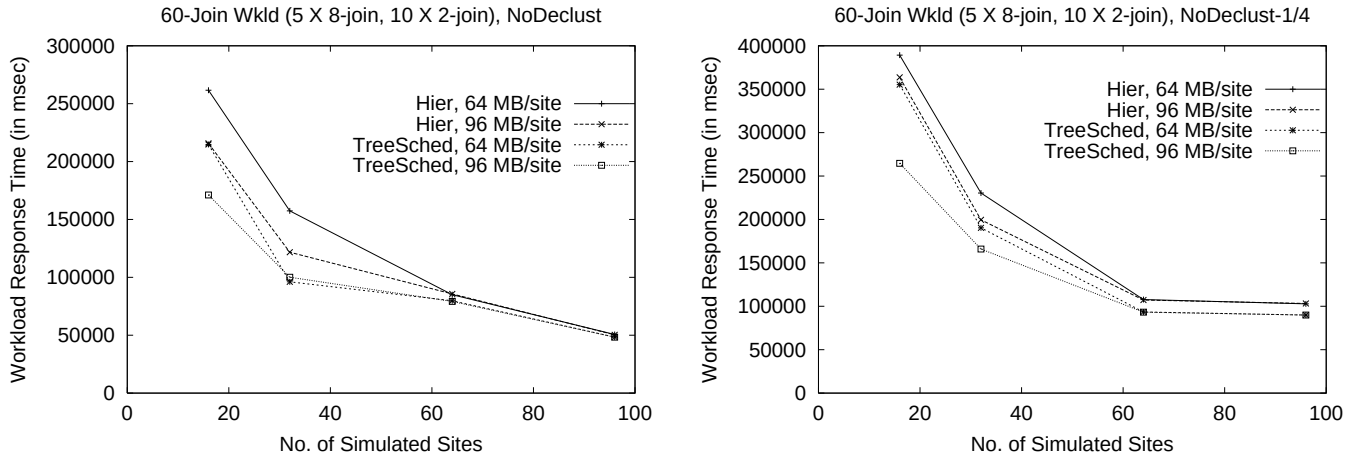

Figure 13: Performance of TREESCHED and HIER for (a) **NoDeclust** and (b) **NoDeclust-1/4**. ($f = 0.002$, $\lambda = 0.2$)

The performance of the three scheduling algorithms under our fully-declustered data placement policy **Declust** is depicted in Figure 14. As our results demonstrate, TREESCHED, HIER, and ZETA perform almost identically under **Declust**. This was something we expected for the following reasons. First, given that all base relation scans for a pipeline occur concurrently at all sites and all result stores are again defaulted to execute on all sites, TREESCHED really has no opportunities to exploit its multi-dimensional cost model in order to balance CPU and I/O processing across system sites. Second, since the startup and synchronization overhead of parallelism is rather small in our simulator configuration, cloning every join across all sites does not penalize performance, at least for the range of system sizes considered in our experiments. Thus, both the one-dimensional HIER and the naive ZETA perform well, managing to consistently stay very close to TREESCHED. (The slight degradation in the performance of ZETA for small configurations is due to the fact that some pipelines are executed in multiple passes, incurring intermediate I/O.) Note that, earlier work (e.g., [Meh94, MD95, MD97]) has shown that the naive strategy of cloning every query operator across all sites can result in very poor execution schedules for system configurations with either higher parallelization overheads or larger numbers of sites than our assumed simulator configuration. In fact, we have also experimented with simulator configurations with lower CPU speeds and we have found that, in many situations, the response times for ZETA actually *increase* as the number of sites exceeds 32.

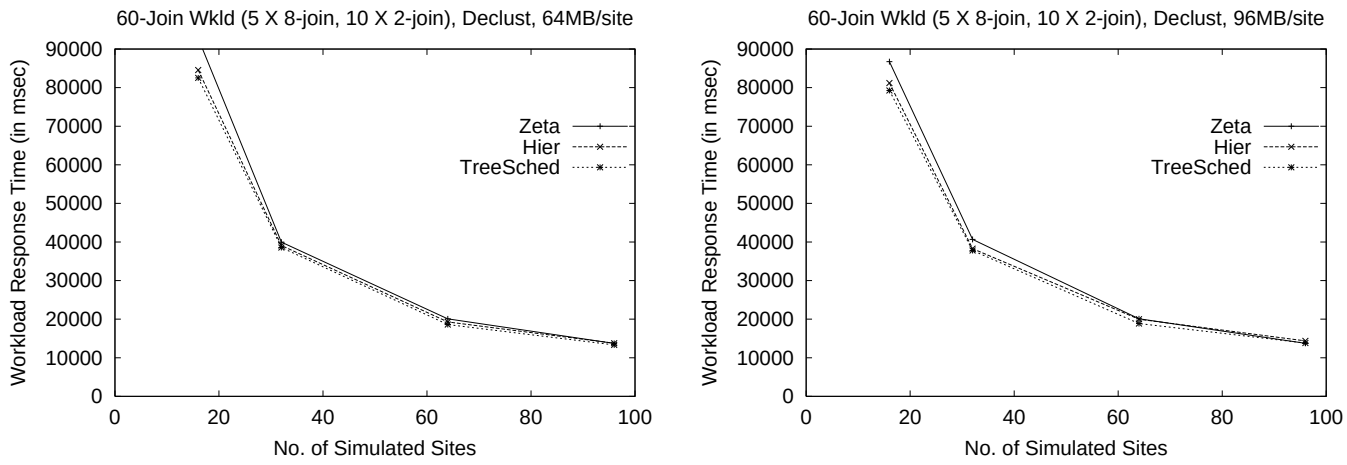

Figure 14: Performance of TreeSched, Hier, and Zeta for **Declust**. ($f = 0.002$, $\lambda = 0.2$)

Figure 15 depicts the response times of the schedules produced by TreeSched, Hier, and Zeta under our **Declust-1/4** and **Random** data placement policies. Once again, both TreeSched and Hier beat Zeta by a wide margin, since they are able to balance the load across the system without being restricted by data placement decisions. Furthermore, TreeSched is also able to exploit its multi-dimensional scheduling model to better balance CPU and disk requirements at sites and provide significant response time gains of up to 30% over the one-dimensional Hier scheduler. Note that, as with the "no-declustering" results, our multi-dimensional scheduler gets its most convincing victories over Hier with resource-limited configurations (e.g., 16 sites with 64 MBytes/site). It is for such resource-limited scenarios that effective resource scheduling becomes a major concern for query processing and optimization.

Finally, the performance of the three scheduling algorithms under our **QueryBasedDeclust** data placement policy is shown in Figure 16. In this particular experiment, we activated the "maximum memory" flag for the Zeta algorithm which ensures that a join will only be executed with its maximum memory allocation. We found that this really helped the response time performance of Zeta under **QueryBasedDeclust**. Still, both TreeSched and Hier manage to outperform Zeta by a significant margin across our entire system size range. A main reason for this phenomenon is that, by the **QueryBasedDeclust** data placement rule, Zeta essentially decides the degree of parallelism for a join by looking only at the `build` relation, which is not always a good predictor for the amount of work that needs to be done. For example, consider a very small `build` relation that is joined with a very large `probe` input. Zeta will probably assign only one site to that relation, which means that it will very likely be a bottleneck during the `probe` phase of the pipeline. On the other hand, both TreeSched and Hier consider both the memory *and* work requirements of a join when making their parallelization decisions. TreeSched is once again able to exploit its multi-dimensional model to better balance CPU and I/O processing and get some performance gains over the one-dimensional Hier schedules. Nevertheless, the benefits in this case are somewhat limited (up to 15%) since, by virtue of the **QueryBasedDeclust** data placement policy, the I/O workload has already been somewhat "uniformly" distributed across the system sites.

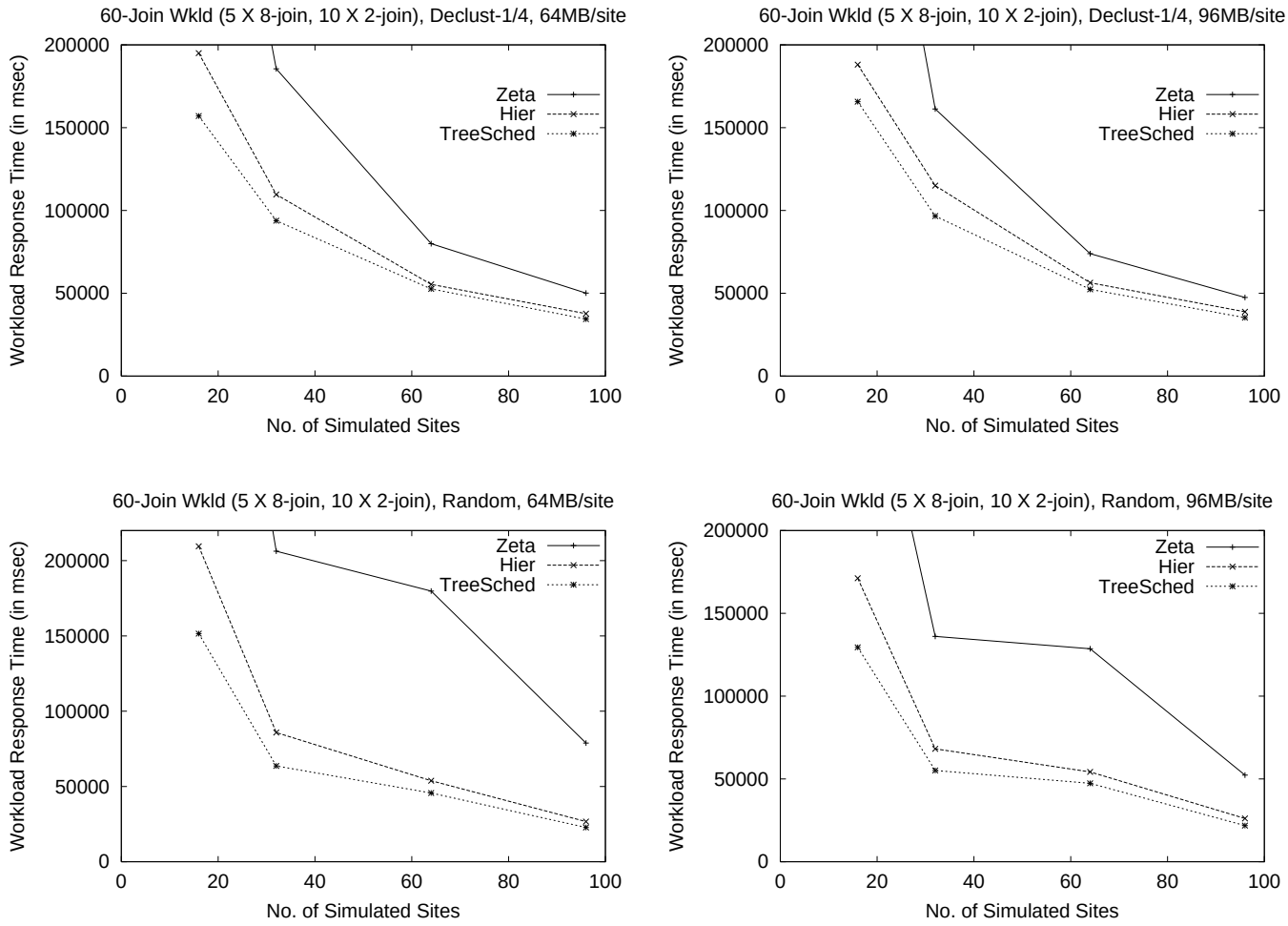

Figure 15: Performance of TreeSched, Hier, and Zeta for (a) **Declust-1/4** (top half), and (b) **Random** (bottom half). ($f = 0.002$, $\lambda = 0.2$)

6 Parallel Query Optimization

In this section, we study the implications of the analytical and experimental results presented in this paper for the open problem of designing efficient cost models for parallel query optimization [DG92].

As noted in Section 1, the use of *response time* as optimization metric implies that a parallel query optimizer cannot afford to ignore resource scheduling during the optimization process. Prior work has demonstrated that a *two-phase* approach [HS91, Hon92] using the traditional work metric during the plan generation phase often results in plans that are inherently sequential and often unable to exploit the available parallelism [JPS93, BFG+95]. On the other hand, using a detailed resource scheduling model during plan generation (as advocated by the *one-phase* approach [SE93, JPS93, LVZ93]) can have a tremendous impact on optimizer complexity and optimization cost. For example, a Dynamic Programming (DP) algorithm must use much stricter pruning criteria that account for the use of system resources [GHK92, LVZ93]. This leads to a combinatorial explosion in the state that must be maintained while building the DP tree, rendering the algorithm impractical even for small query sizes.

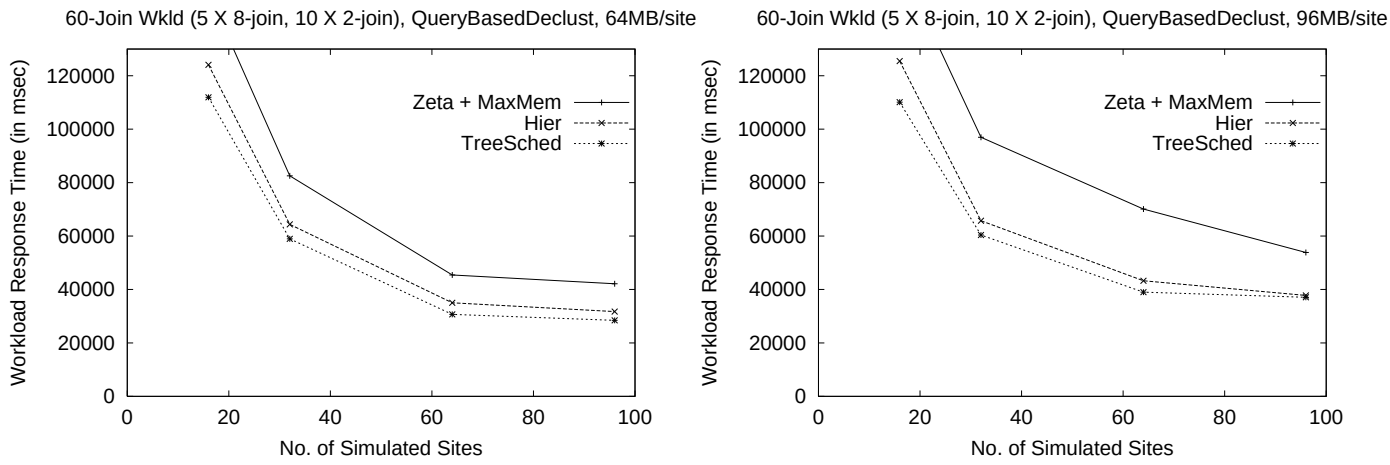

Figure 16: Performance of TreeSched, Hier, and Zeta for **QueryBasedDeclust**. ($f = 0.002$, $\lambda = 0.2$)

As in centralized query optimization, the role of the optimizer cost model is to provide an abstraction of the underlying execution system. In this respect, the one- and two-phase approaches lie at the two different ends of a spectrum, incorporating either detailed knowledge (one-phase) or no knowledge (two-phase) of the parallel execution environment in the optimizer cost metric. The goal is to devise cost metrics that are more realistic than total resource consumption, in the sense that they are cognizant of the available parallelism, and at the same time are sufficiently efficient to keep the optimization process tractable.

In earlier work, Ganguly et al. [GGS96] suggested the use of a novel scalar cost metric for parallel query optimization. Their metric was defined as the maximum of two "bulk parameters" of a parallel query plan, namely the critical path length of the plan tree and the average work per site. Although the model used in the work of Ganguly et al. was one-dimensional, it is clear that the "critical path length" corresponds to the maximum, over all root-to-leaf paths, sum of T^{max}'s in the task tree, whereas the "average work" corresponds to $\frac{l(S^W)}{P}$ with S being all operator clones in the plan.

Based on our analytical and experimental results, there clearly exists a third parameter, namely the *average volume per site* $\frac{l(S^{TV})}{P}$ that is an essential component of query plan quality. The importance of this third parameter stems from the fact that it is the only one capturing the constraints on parallel execution that derive from ss (i.e., memory) resources.

We believe that the triple *(critical path, average work, average volume)* captures all the crucial aspects characterizing the expected response time of a parallel query execution plan. Consequently, we feel that these three components can provide the basis for an efficient and accurate cost model for parallel query optimizers. They can be used exactly as Ganguly et al. [GGS96] suggested, i.e., combined through a max{} function to produce a scalar metric. Alternatively, they can be used as a 3-dimensional vector together with a 3-dimensional "less than" to compare plans (e.g., to prune the search space in a DP-based parallel optimizer [GHK92]). Clearly, using only three dimensions turns the Partial Order DP (PODP) approach of Ganguly et al. [GHK92] into a feasible and efficient paradigm for DP-based parallel query optimization.

7 Related Work

The problem of scheduling complex query plans on parallel machines has attracted a lot of attention from the database research community. Hasan and Motwani [HM94] study the tradeoff between pipelined parallelism and its communication overhead and develop near-optimal heuristics for scheduling a star or a chain of pipelined relational operators on a multiprocessor architecture. Chekuri et al. [CHM95] extend these results to arbitrary pipelined operator trees. The heuristics proposed in these papers ignore both independent and partitioned parallelism. Ganguly and Wang [GW93] describe the design of a parallelizing scheduler for a tree of coarse grain operators. Based on a *one-dimensional* model of query operator costs (that combines the operators' CPU and disk work in a scalar "execution time" metric), the authors show their scheduler to be near-optimal for a limited space of query plans (i.e., left-deep join trees with a single materialization point in any right subtree). Ganguly et al. [GGW95] obtain similar results for the problem of partitioning independent pipelines without the coarse granularity restriction. The benefits of resource sharing and the multi-dimensionality of query operators are not addressed in these papers. Furthermore, no experimental results are reported. Lo et al. [LCRY93] develop optimal schemes for assigning processors to the stages of a pipeline of hash-joins in a shared-disk environment. Their schemes are based on a *two-phase minimax* formulation of the problem that ignores communication costs and prevents processor sharing among stages. Moreover, no methods are proposed for handling multiple join pipelines (i.e., independent parallelism).

With the exception of the papers mentioned above, most efforts are experimental in nature and offer no theoretical justification for the algorithms that they propose. In addition, many proposals have simplified the scheduling issues by ignoring independent (bushy tree) parallelism; these include the right-deep trees of Schneider [Sch90a] and the segmented right-deep trees of Chen et al. [CLYY92]. Nevertheless, the advantages offered by such parallelism, especially for large queries, have been demonstrated in prior research [CYW92].

Tan and Lu [TL93] and Niccum et al. [NSHL95] consider the general problem of scheduling bushy join plans on parallel machines exploiting all forms of intra-query parallelism and suggest heuristic methods of splitting the bushy plan into non-overlapping *shelves* of concurrent joins. For the same problem, Hsiao et al. [HCY94] propose a processor allocation scheme based on the concept of *synchronous execution time*: the set of processors allotted to a parent join pipeline are recursively partitioned among its subtrees in such a way that those subtrees can be completed at approximately the same time. For deep execution plans, there exists a point beyond which further partitioning is detrimental or even impossible, and serialization must be employed for better performance. Wolf et al. [WTCY94, WTCY95] present a one-dimensional hierarchical algorithm for scheduling multiple parallel queries on a set of processors. Their main idea is to collapse each query plan to a single "large" parallel job and then apply the known results for independent jobs. This has the serious drawback that some obvious, critical co-scheduling may be lost. For example, although it may be highly desirable to combine CPU-bound and IO-bound tasks from different plans [Hon92], this may not be possible after the collapse. Wilschut et al. [WFA95] present a comparative performance evaluation of various multi-join execution strategies on the PRISMA/DB parallel main-memory database system. Mehta and DeWitt [MD95] and Rahm and Marek [RM95] present experimental evaluations of various heuristic strategies for determining the degree of intra-operation parallelism and assigning processors in shared-nothing DBMSs. Both of these papers avoid dealing with complex query scheduling issues by assuming workloads consisting of simple binary joins and/or OLTP transactions. Bouganim et al. [BFV96] propose methods for optimizing load-balancing on each site of a hierarchical architecture at run-time so that inter-

site data transfers are minimized. In their model, the optimizer still has to determine the assignment of operators to sites and the run-time environment has to make up for optimizer inaccuracies. The issue of how the high-level mapping should be done at the optimizer is not addressed.

A common characteristic of all approaches described above is that they consider a one-dimensional model of resource allocation based on a scalar cost metric (e.g., "work"), which ignores any possibilities for effective resource sharing among concurrent operations. Perhaps the only exception is Hong's method for exploiting independent parallelism in the XPRS shared-memory database system [Hon92]. His approach is based on dynamically balancing resource use between one I/O-bound and one CPU-bound operator pipeline to ensure that the system always executes at its IO-CPU balance point. However, the substantial cost of communication renders such a scheduling method impractical for shared-nothing architectures. In more recent work, Nippl and Mitschang [NM98] discuss the design of TOPAZ, a rule-driven, multi-phase parallelizer with a cost model that takes into account CPU, disk, and communication costs, as well as memory usage. However, the target architecture of the TOPAZ parallelizer is based on the shared-disk paradigm and the proposed strategies are primarily concerned with determining the degree of parallelism for operators and not with mapping operator clones to distributed resource sites.

Moving away from the database field, there is a significant body of work on parallel task scheduling in the field of deterministic scheduling theory. Since the problem is $\mathcal{NP}$-hard in the strong sense [DL89], research efforts have concentrated on providing fast heuristics with provable worst case bounds on the suboptimality of the solution. However, scheduling query plans on shared-nothing parallel architectures requires a significantly richer model of parallelization than what is assumed in the classical [Gra66, GG75, GLLRK79] or more recent [BB90, BB91, KM92, TWY92, WC92, ST99, CM96] efforts in that field. To the best of our knowledge, there have been no theoretical results in the literature on parallel task scheduling that consider multiple TS system resources and explore sharing of such resources among concurrent tasks, or study the implications of pipelined parallelism and data communication costs. Perhaps most importantly, even the very recent results of Chakrabarti and Muthukrishnan [CM96] and Shachnai and Turek [ST99] on multi-resource scheduling are based on the assumption that all resources are *globally accessible* to all tasks. In contrast, our target architectures are characterized by a physical distribution of resource units and an *affinity* of system resources to sites: an operation scheduled at a particular site can only make use of the resources locally available to that operation. To the best of our knowledge, there are no previous theoretical results on multi-resource scheduling in this context.

Finally, it is worth noting that the two resource classes (i.e., TS and SS) considered in this paper have been identified in prior work, e.g., the "stretchable" and "non-stretchable" resources of Pirahesh et al. [PMC+90] and Ganguly et al. [GHK92]. The idea of lower bounding the makespan of a schedule based on a resource-time product (like our "volume" bound) dates back to the seventies and the pioneering work of Garey and Graham on multi-resource scheduling [GG75]. Yu and Cornell [YC93] have used the memory-time product of a join operator to define its "return on consumption" which, as they demonstrate, can be used to effectively allocate memory among multiple competing single-join queries. Other work has also suggested heuristic strategies for allocating memory statically or dynamically to complex query plans [BKV98, ND98]. We should stress, however, that the general problem of scheduling operator graphs on a parallel system with both TS and SS resources has not been addressed in prior work on databases or deterministic scheduling theory.

8 Conclusions

The problem of scheduling complex queries in shared-nothing parallel database systems of multiple time-shared and space-shared resources has been open for a long time both within the database field and the deterministic scheduling theory field. Despite the importance of such architectures in practice, the difficulties of the problem have led researchers in making various assumptions and simplifications that are not realistic. In this paper, we have provided what we believe is the first comprehensive formal approach to the problem. We have established a model of resource usage that allows the scheduler to explore the possibilities for concurrent operations sharing both TS and SS resources and quantify the effects of this sharing on the parallel execution time. The inclusion of both types of resources has given rise to interesting tradeoffs with respect to the degree of partitioned parallelism, which are nicely exposed within our analytical models and results, and for which we have provided some effective resolutions. We have provided efficient, near-optimal heuristic algorithms for query scheduling in such parallel environments, paying special attention to various constraints that arise from the existence of SS resources, including the co-scheduling requirements of pipelined operator execution, which has been the most challenging to resolve. Our set of results apply to all types of query plans and even sets of plans that are either provided all at the beginning or arrive dynamically for scheduling. As a side-effect of our effort, we have identified the importance of a parameter that captures one aspect of parallel query execution cost, which should play an important role in obtaining realistic cost models for parallel query optimization. Finally, we have presented a set of experimental results from the implementation of our scheduling algorithms on top of a detailed simulation environment for shared-nothing database systems based on the Gamma parallel database machine. These results have verified the effectiveness of our scheduling algorithms in a realistic system setting.

Our effort suggests several directions for future research. First, it is a challenging problem to extend our multi-dimensional scheduling algorithms and results to the more general *malleable* scheduling scenario, in which the scheduler is allowed to trade off some types of resources for others. For example, the number of pages reserved for a hash-join operator can be anywhere within a range of possible allocations with smaller allocations typically implying more disk I/O. Scheduling multiple such malleable operators is a very hard problem. Second, we would like to investigate the effect of our results on parallel query optimization and see if they lead to efficient and accurate optimizers. Finally, given the generality of our scheduling framework, it would be interesting to investigate its applicability to other similar environments, such as multimedia storage servers. These questions form the basis of our current and future research.

References

[BB90] Krishna P. Belkhale and Prithviraj Banerjee. "Approximate Algorithms for the Partitionable Independent Task Scheduling Problem". In *Proceedings of the 1990 International Conference on Parallel Processing*, pages I72–I75, August 1990.

[BB91] Krishna P. Belkhale and Prithviraj Banerjee. "A Scheduling Algorithm for Parallelizable Dependent Tasks". In *Proceedings of the Fifth International Parallel Processing Symposium*, pages 500–506, 1991.

[BFG+95] C.K. Baru, G. Fecteau, A. Goyal, H. Hsiao, A. Jhingran, S. Padmanabhan, G.P. Copeland, and W.G. Wilson. "DB2 Parallel Edition". *IBM Systems Journal*, 34(2):292–322, 1995.

[BFMV00] Luc Bouganim, Françoise Fabret, C. Mohan, and Patrick Valduriez. "Dynamic Query Scheduling in Data Integration Systems". In *Proceedings of the Sixteenth International Conference on Data Engineering*, pages 425–434, San Diego, California, March 2000.

[BFV96] Luc Bouganim, Daniela Florescu, and Patrick Valduriez. "Dynamic Load Balancing in Hierarchical Parallel Database Systems". In *Proceedings of the 22nd International Conference on Very Large Data Bases*, pages 436–447, Mumbai(Bombay), India, September 1996.

[BKV98] Luc Bouganim, Olga Kapitskaia, and Patrick Valduriez. "Memory-Adaptive Scheduling for Large Query Execution". In *Proceedings of the Seventh International Conference on Information and Knowledge Management*, pages 105–115, Bethesda, Maryland, November 1998.

[Bro94] Kurt Brown. "PRPL: A Database Workload Specification Language (Version 1.4)". Unpublished Manuscript, March 1994.

[BS83] Brenda S. Baker and Jerald S. Schwarz. "Shelf Algorithms for Two-Dimensional Packing Problems". *SIAM Journal on Computing*, 12(3):508–525, August 1983.

[CGJ84] E.G. Coffman, Jr., M.R. Garey, and D.S. Johnson. "Approximation Algorithms for Bin-Packing – An Updated Survey". In *"Algorithm Design for Computing System Design"*, pages 49–106. Springer-Verlag, New York, 1984.

[CGJT80] E.G. Coffman, Jr., M.R. Garey, D.S. Johnson, and R.E. Tarjan. "Performance Bounds for Level-Oriented Two-Dimensional Packing Algorithms". *SIAM Journal on Computing*, 9(4):808–826, November 1980.

[CHM95] Chandra Chekuri, Waqar Hasan, and Rajeev Motwani. "Scheduling Problems in Parallel Query Optimization". In *Proceedings of the Fourteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 255–265, San Jose, California, May 1995.

[CLYY92] Ming-Syan Chen, Ming-Ling Lo, Philip S. Yu, and Honesty C. Young. "Using Segmented Right-Deep Trees for the Execution of Pipelined Hash Joins". In *Proceedings of the Eighteenth International Conference on Very Large Data Bases*, pages 15–26, Vancouver, Canada, August 1992.

[CM96] Soumen Chakrabarti and S. Muthukrishnan. "Resource scheduling for parallel database and scientific applications". In *Proceedings of the Eighth Annual ACM Symposium on Parallel Algorithms and Architectures*, pages 329–335, Padua, Italy, June 1996.

[CYW92] Ming-Syan Chen, Philip S. Yu, and Kun-Lung Wu. "Scheduling and Processor Allocation for Parallel Execution of Multi-Join Queries". In *Proceedings of the Eighth International Conference on Data Engineering*, pages 58–67, Phoenix, Arizona, February 1992.

[DG92] David J. DeWitt and Jim Gray. "Parallel Database Systems: The Future of High Performance Database Database Systems". *Communications of the ACM*, 35(6):85–98, June 1992.

[DGS+90] David J. DeWitt, Shahram Ghandeharizadeh, Donovan A. Schneider, Allan Bricker, Hui-I Hsiao, and Rick Rasmussen. "The Gamma Database Machine Project". *IEEE Transactions on Knowledge and Data Engineering*, 2(1):44–62, March 1990.

[DL89] Jianzhong Du and Joseph Y-T. Leung. "Complexity of Scheduling Parallel Task Systems". *SIAM Journal on Discrete Mathematics*, 2(4):473–487, November 1989.

[EGH95] Susanne Englert, Ray Glasstone, and Waqar Hasan. "Parallelism and its Price: A Case Study of NonStop SQL/MP". *ACM SIGMOD Record*, 24(4):61–71, December 1995.

[GG75] M.R. Garey and R.L. Graham. "Bounds for Multiprocessor Scheduling with Resource Constraints". *SIAM Journal on Computing*, 4(2):187–200, June 1975.

[GGS96] Sumit Ganguly, Akshay Goel, and Avi Silberschatz. "Efficient and Accurate Cost Models for Parallel Query Optimization". In *Proceedings of the Fifteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, Montreal, Quebec, June 1996.

[GGW95] Sumit Ganguly, Apostolos Gerasoulis, and Weining Wang. "Partitioning Pipelines with Communication Costs". In *Proceedings of the 6th International Conference on Information Systems and Data Management (CISMOD'95)*, pages 302–320, Bombay, India, November 1995.

[GHK92] Sumit Ganguly, Waqar Hasan, and Ravi Krishnamurthy. "Query Optimization for Parallel Execution". In *Proceedings of the 1992 ACM SIGMOD International Conference on Management of Data*, pages 9–18, San Diego, California, June 1992.

[GI96] Minos N. Garofalakis and Yannis E. Ioannidis. "Multi-dimensional Resource Scheduling for Parallel Queries". In *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, pages 365–376, Montreal, Quebec, June 1996.

[GI97] Minos N. Garofalakis and Yannis E. Ioannidis. "Parallel Query Scheduling and Optimization with Time- and Space-Shared Resources". In *Proceedings of the 23rd International Conference on Very Large Data Bases*, pages 296–305, Athens, Greece, August 1997.

[GIÖ98] Minos N. Garofalakis, Yannis E. Ioannidis, and Banu Özden. "Resource Scheduling for Composite Multimedia Objects". In *Proceedings of the 24th International Conference on Very Large Data Bases*, New York, USA, August 1998.

[GJ79] M.R. Garey and D.S. Johnson. *"Computers and Intractability: A Guide to the Theory of NP-Completeness"*. W.H. Freeman, 1979.

[GLLRK79] R.L. Graham, E.L. Lawler, J.K. Lenstra, and A.H.G. Rinnooy Kan. "Optimization and Approximation in Deterministic Sequencing and Scheduling: A Survey". *Annals of Discrete Mathematics*, 5:287–326, 1979.

[GMSY93] Shahram Ghandeharizadeh, Robert R. Meyer, Gary L. Schultz, and Jonathan Yackel. "Optimal Balanced Assignments and a Parallel Database Application". *ORSA Journal on Computing*, 5(2):151–167, Spring 1993.

[Gra66] R.L. Graham. "Bounds for Certain Multiprocessing Anomalies". *The Bell System Technical Journal*, 45:1563–1581, November 1966.

[Gra69] R.L. Graham. "Bounds on Multiprocessing Timing Anomalies". *SIAM Journal on Computing*, 17(2):416–429, March 1969.

[Gra93] Goetz Graefe. "Query Evaluation Techniques for Large Databases". *ACM Computing Surveys*, 25(2):73–170, June 1993.

[GW93] Sumit Ganguly and Weining Wang. "Optimizing Queries for Coarse Grain Parallelism". Technical Report LCSR-TR-218, Rutgers University, Dept. of Computer Sciences, October 1993.

[Has95] Waqar Hasan. *"Optimization of SQL Queries for Parallel Machines"*. PhD thesis, Stanford University, December 1995.

[HCL+90] Laura M. Haas, Walter Chang, Guy M. Lohman, John McPherson, Paul F. Wilms, George Lapis, Bruce G. Lindsay, Hamid Pirahesh, Michael J. Carey, and Eugene J. Shekita. "Starburst Mid-Flight: As the Dust Clears". *IEEE Transactions on Knowledge and Data Engineering*, 2(1):143–160, March 1990.

[HCY94] Hui-I Hsiao, Ming-Syan Chen, and Philip S. Yu. "On Parallel Execution of Multiple Pipelined Hash Joins". In *Proceedings of the 1994 ACM SIGMOD International Conference on Management of Data*, pages 185–196, Minneapolis, Minnesota, May 1994.

[HFV96] Waqar Hasan, Daniela Florescu, and Patrick Valduriez. "Open Issues in Parallel Query Optimization". *ACM SIGMOD Record*, 25(3):28–33, September 1996.

[HM94] Waqar Hasan and Rajeev Motwani. "Optimization Algorithms for Exploiting the Parallelism-Communication Tradeoff in Pipelined Parallelism". In *Proceedings of the 20th International Conference on Very Large Data Bases*, pages 36–47, Santiago, Chile, August 1994.

[Hon92] Wei Hong. "Exploiting Inter-Operation Parallelism in XPRS". In *Proceedings of the 1992 ACM SIGMOD International Conference on Management of Data*, pages 19–28, San Diego, California, June 1992.

[HS91] Wei Hong and Michael Stonebraker. "Optimization of Parallel Query Execution Plans in XPRS". In *Proceedings of the First International Conference on Parallel and Distributed Information Systems*, Miami Beach, Florida, December 1991.

[HS93] Wei Hong and Michael Stonebraker. "Optimization of Parallel Query Execution Plans in XPRS". *Distributed and Parallel Databases*, 1:9–32, 1993.

[Ioa93] Yannis E. Ioannidis. "Universality of Serial Histograms". In *Proceedings of the Nineteenth International Conference on Very Large Data Bases*, pages 256–267, Dublin, Ireland, August 1993.

[JMP97] Anant Jhingran, Timothy Malkemus, and Sriram Padmanabhan. "Query Optimization in DB2 Parallel Edition". *IEEE Data Engineering Bulletin*, 20(2):27–34, June 1997. (Special Issue on Commercial Parallel Systems).

[JPS93] Anant Jhingran, Sriram Padmanabhan, and Ambuj Shatdal. "Join Query Optimization in Parallel Database Systems". In *Proceedings of the IEEE Workshop on Advances in Parallel and Distributed Systems*, 1993.

[KM92] Ramesh Krishnamurti and Eva Ma. "An Approximation Algorithm for Scheduling Tasks on Varying Partition Sizes in Partitionable Multiprocessor Systems". *IEEE Transactions on Computers*, 41(12):1572–1579, December 1992.

[LCRY93] Ming-Ling Lo, Ming-Syan Chen, C.V. Ravishankar, and Philip S. Yu. "On Optimal Processor Allocation to Support Pipelined Hash Joins". In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 69–78, Washington, D.C., June 1993.

[LVZ93] Rosana S.G. Lanzelotte, Patrick Valduriez, and Mohamed Zaït. "On the Effectiveness of Optimization Search Strategies for Parallel Execution Spaces". In *Proceedings of the Nineteenth International Conference on Very Large Data Bases*, pages 493–504, Dublin, Ireland, August 1993.

[MD95] Manish Mehta and David J. DeWitt. "Managing Intra-operator Parallelism in Parallel Database Systems". In *Proceedings of the 21st International Conference on Very Large Data Bases*, pages 382–394, Zurich, Switzerland, September 1995.

[MD97] Manish Mehta and David J. DeWitt. "Data Placement in Shared-Nothing Parallel Database Systems". *The VLDB Journal*, 6(1):53–72, January 1997.

[Meh94] Manish Mehta. *"Resource Allocation for Parallel Shared-Nothing Database Systems"*. PhD thesis, University of Wisconsin-Madison, 1994.

[ND98] Biswadeep Nag and David J. DeWitt. "Memory Allocation Strategies for Complex Decision Support Queries". In *Proceedings of the Seventh International Conference on Information and Knowledge Management*, Bethesda, Maryland, November 1998.

[NM98] Clara Nippl and Bernhard Mitschang. "TOPAZ: a Cost-Based, Rule-Driven, Multi-Phase Parallelizer". In *Proceedings of the 24th International Conference on Very Large Data Bases*, New York, USA, August 1998.

[NSHL95] Thomas M. Niccum, Jaideep Srivastava, Bhaskar Himatsingka, and Jianzhong Li. "Query Optimization and Processing in Parallel Databases". *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 22:259–287, 1995.

[NZT96] Michael G. Norman, Thomas Zurek, and Peter Thanisch. "Much Ado About Shared-Nothing". *ACM SIGMOD Record*, 25(3):16–21, September 1996.

[ÖRS96] Banu Özden, Rajeev Rastogi, and Avi Silberschatz. "Disk Striping in Video Server Environments". In *Proceedings of the 1996 International Conference on Multimedia Computing and Systems*, Hiroshima, Japan, June 1996.

[ÖRS97] Banu Özden, Rajeev Rastogi, and Avi Silberschatz. "Multimedia Support for Databases". In *Proceedings of the Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, Tucson, Arizona, May 1997.

[PI96] Viswanath Poosala and Yannis E. Ioannidis. "Estimation of Query-Result Distribution and its Application in Parallel-Join Load Balancing". In *Proceedings of the 22nd International Conference on Very Large Data Bases*, pages 448–459, Mumbai(Bombay), India, September 1996.

[PMC+90] Hamid Pirahesh, C. Mohan, Josephine Cheng, T.S. Liu, and Pat Selinger. "Parallelism in Relational Data Base Systems: Architectural Issues and Design Approaches". In *Proceedings of the Second International Symposium on Databases in Parallel and Distributed Systems*, pages 4–29, Dublin, Ireland, July 1990.

[RM95] Erhard Rahm and Robert Marek. "Dynamic Multi-Resource Load Balancing in Parallel Database Systems". In *Proceedings of the 21st International Conference on Very Large Data Bases*, pages 395–406, Zurich, Switzerland, September 1995.

[SAC+79] P. Selinger, M.M. Astrahan, D.D. Chamberlin, R.A. Lorie, and T.G. Price. "Access Path Selection in a Relational Database Management System". In *Proceedings of the 1979 ACM SIGMOD International Conference on Management of Data*, pages 23–34, Boston, Massachusetts, June 1979.

[Sch90a] Donovan A. Schneider. *"Complex Query Processing in Multiprocessor Database Machines"*. PhD thesis, University of Wisconsin-Madison, September 1990.

[Sch90b] Herb Schwetman. "CSIM User's Guide". Technical Report ACT-126-90, MCC, Austin, Texas, March 1990.

[SE93] Jaideep Srivastava and Gary Elsesser. "Optimizing Multi-Join Queries in Parallel Relational Databases". In *Proceedings of the Second International Conference on Parallel and Distributed Information Systems*, pages 84–92, San Diego, California, January 1993.

[Sea] Seagate Technology. Cheetah 36 (ST136403FC) Technical Specifications. (`http://www.seagate.com/`).

[SG98] Abraham Silberschatz and Peter Galvin. *"Operating System Concepts"*. Addison-Wesley Publishing Company, 1998. (Fifth Edition).

[Sha86] Leonard D. Shapiro. "Join Processing in Database Systems with Large Main Memories". *ACM Transactions on Database Systems*, 11(3):239–264, September 1986.

[ST99] Hadas Shachnai and John J. Turek. "Multiresource malleable task scheduling to minimize response time". *Information Processing Letters*, 70(1):47–52, April 1999.

[Sto86] Michael Stonebraker. "The Case for Shared Nothing". *IEEE Data Engineering Bulletin*, 9(1):4–9, March 1986.

[SYT93] Eugene J. Shekita, Honesty C. Young, and Kian-Lee Tan. "Multi-Join Optimization for Symmetric Multiprocessors". In *Proceedings of the Nineteenth International Conference on Very Large Data Bases*, pages 479–492, Dublin, Ireland, August 1993.

[TL93] Kian-Lee Tan and Hongjun Lu. "On Resource Scheduling of Multi-join Queries in Parallel Database Systems". *Information Processing Letters*, 48:189–195, 1993.

[TWPY92] John Turek, Joel L. Wolf, Krishna R. Pattipati, and Philip S. Yu. "Scheduling Parallelizable Tasks: Putting it All on the Shelf". In *Proceedings of the 1992 ACM SIGMETRICS Conference on Measurement & Modeling of Computer Systems*, pages 225–236, Newport, Rhode Island, June 1992.

[TWY92] John Turek, Joel L. Wolf, and Philip S. Yu. "Approximate Algorithms for Scheduling Parallelizable Tasks". In *Proceedings of the Fourth Annual ACM Symposium on Parallel Algorithms and Architectures*, pages 323–332, San Diego, California, June 1992.

[Val93] Patrick Valduriez. "Parallel Database Systems: Open Problems and New Issues". *Distributed and Parallel Databases*, 1:137–165, 1993.

[WC92] Qingzhou Wang and Kam Hoi Cheng. "A Heuristic of Scheduling Parallel Tasks and its Analysis". *SIAM Journal on Computing*, 21(2):281–294, April 1992.

[WFA92] Annita N. Wilschut, Jan Flokstra, and Peter M.G. Apers. "Parallelism in a Main-Memory DBMS: The Performance of PRISMA/DB". In *Proceedings of the Eighteenth International Conference on Very Large Data Bases*, pages 521–532, Vancouver, Canada, August 1992.

[WFA95] Annita N. Wilschut, Jan Flokstra, and Peter M.G. Apers. "Parallel Evaluation of Multi-join Queries". In *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data*, pages 115–126, San Jose, California, May 1995.

[WTCY94] Joel L. Wolf, John Turek, Ming-Syan Chen, and Philip S. Yu. "Scheduling Multiple Queries on a Parallel Machine". In *Proceedings of the 1994 ACM SIGMETRICS Conference on Measurement & Modeling of Computer Systems*, pages 45–55, Nashville, Tennessee, May 1994.

[WTCY95] Joel L. Wolf, John Turek, Ming-Syan Chen, and Philip S. Yu. "A Hierarchical Approach to Parallel Multiquery Scheduling". *IEEE Transactions on Parallel and Distributed Systems*, 6(6):578–590, June 1995.

[YC93] Philip S. Yu and Douglas W. Cornell. "Buffer Management Based on Return on Consumption in a Multi-Query Environment". *The VLDB Journal*, 2(1):1–37, January 1993.

A Proofs of Theoretical Results

Proof of Lemma 4.1: Given our earlier results for the case of only TS resources [GI96], we only need to establish the result concerning the new "volume" term in the max function. Let T_{ij}, $\overline{W}_{ij}$, and $\overline{V}_{ij}$ denote, respectively, the stand-alone time, work vector, and demand vector of the j-th clone of $\mathtt{op}_i$ ($j = 1, \ldots, N_i$). Note that the j-th clone of $\mathtt{op}_i$ will require resource fractions $\overline{V}_{ij}$ for *at least* T_{ij} time. (It may be longer if the clone experiences contention on some preemptable resource(s).) Thus, for *any* schedule SCHED, the total resource-time product for the parallel execution must exceed these requirements for all non-preemptable resources. That is, the parallel execution time of SCHED on P sites should satisfy the following (componentwise) inequality:

$$T^{par}(\text{SCHED}, P)\, P\, \overline{1} \geq_s \sum_{i,j} T_{ij}\overline{V}_{ij},$$

where $\overline{1}$ denotes an s-dimensional vector of 1's. Taking this inequality for the maximum component of the right hand side gives the desired result. ∎

Proof of Theorem 4.1: Let n_i denote the number of compatible clone subsets in site B_i and let S_{ij} be the j-th such subset. Define S_{ij}^V, S_{ij}^W, S_{ij}^{TV}, and $T^{max}(S_{ij})$ in the usual manner (see Section 4.1). Consider any two consecutive compatible subsets S_{ij} and $S_{i,j+1}$ in site B_i. Let $c_m = (T_m, \overline{W}_m, \overline{V}_m)$ be the first clone placed in $S_{i,j+1}$. By the operation of the algorithm we know that $l(S_{ij}^V) + l(\overline{V}_m) > 1$. Further, by the order of clone placement we know that for all clones c_k placed in S_{ij} we will have $T_k \geq T_m = T^{max}(S_{i,j+1})$. So, considering the total volume packed on the two shelves we have:

$$l(S_{ij}^{TV}) + l(S_{i,j+1}^{TV}) \geq T^{max}(S_{i,j+1}) \cdot l(S_{ij}^V) + T^{max}(S_{i,j+1}) \cdot l(\overline{V}_m) > T^{max}(S_{i,j+1}).$$

Thus, taking the sum of T^{max}'s over all compatible subsets in B_i we have:

$$\sum_{j=1}^{n_i} T^{max}(S_{ij}) < T^{max}(S_{i,1}) + \sum_{j=2}^{n_i} l(S_{i,j-1}^{TV}) + \sum_{j=2}^{n_i} l(S_{ij}^{TV}) \leq T^{max}(S) + 2 \cdot \sum_{j=1}^{n_i} l(S_{ij}^{TV}).$$

Consider the placement of the final clone $c_N = (T_N, \overline{W}_N, \overline{V}_N)$. Without loss of generality, assume that the clone is placed in B_1. By the operation of our list scheduling heuristic we know that all the other bins must be of height larger than or equal to the height of B_1. Thus, using the notation of Section 4.2.1, we have

$$P \cdot T^{site}(B_1) \leq \sum_{i,j} T(S_{ij}) = \sum_{i,j} \max\{T^{max}(S_{ij}), l(S_{ij}^W)\} \leq \sum_{i,j} T^{max}(S_{ij}) \; + \; \sum_{i,j} l(S_{ij}^W).$$

And using our previous inequality, we have:

$$P \cdot T^{site}(B_1) < P \cdot T^{max} + 2 \cdot \sum_{i,j} S_{ij}^{TV} + \sum_{i,j} l(S_{ij}^W). \tag{2}$$

The following ancillary lemma establishes an fundamental property of the "vector length" ($l()$) function.

Lemma A.1 *Let S be a set of d-dimensional vectors, and let $\pi = \{S_1, \ldots, S_k\}$ be* **any** *partition of S. Then,*

$$\frac{\sum_{i=1}^{k} l(S_i)}{d} \;\leq\; l(S) \;\leq\; \sum_{i=1}^{k} l(S_i)$$

Proof: The inequality $l(S) \leq \sum_{i=1}^{k} l(S_i)$ is obvious by the definition of $l()$. For the other side of the inequality, let $\overline{v}_i = \sum_{\overline{w} \in S_i} \overline{w}$ for each $i = 1, \ldots, k$. Also, define the function $m(\overline{w}) = \min_{1 \leq j \leq d}\{j : l(\overline{w}) = w[j]\}$, for each d-dimensional vector $\overline{w}$.
Rearranging the sum $\sum_{i=1}^{k} l(S_i)$, we obtain

$$\begin{aligned}
\sum_{i=1}^{k} l(S_i) &= \sum_{i=1}^{k} \max_{1 \leq j \leq d}\{v_i[j]\} \\
&= \sum_{\overline{v}_i : m(\overline{v}_i)=1} \max_{1 \leq j \leq d}\{v_i[j]\} + \ldots + \sum_{\overline{v}_i : m(\overline{v}_i)=d} \max_{1 \leq j \leq d}\{v_i[j]\}
\end{aligned}$$

Observe that, by the definition of $m()$, the $\sum$ and max functions can be legally interchanged in each of the above d terms. Therefore,

$$\begin{aligned}
\sum_{i=1}^{k} l(S_i) &= \max_{1 \leq j \leq d}\{\sum_{\overline{v}_i : m(\overline{v}_i)=1} v_i[j]\} + \ldots + \max_{1 \leq j \leq d}\{\sum_{\overline{v}_i : m(\overline{v}_i)=d} v_i[j]\} \\
&\leq d \max_{1 \leq j \leq d}\{\sum_{i=1}^{k} v_i[j]\} \;=\; d \max_{1 \leq j \leq d}\{\sum_{i=1}^{N} w_i[j]\} \;=\; dl(S)
\end{aligned}$$

Which, of course, implies $l(S) \geq \frac{\sum_{i=1}^{k} l(S_i)}{d}$. ∎

We can now combine Inequality 2 and Lemma A.1 to get:

$$T^{site}(B_1) < T^{max}(S) + 2s \cdot \frac{l(S^{TV})}{P} + d \cdot \frac{l(S^{W})}{P}.$$

Finally, note that the response time of the schedule SCHED obtained by our heuristic will certainly satisfy $T^{par}(\text{SCHED}, P) \leq T^{site}(B_1) + T_N \leq T^{site}(B_1) + T^{max}(S)$. Combining this with the above inequality for $T^{site}(B_1)$ gives the result. ∎

Proof of Lemma 4.2: Assume the claim is false. This implies that there exists at least one SS vector $\overline{v}$ in S^V that cannot fit in any of the $\frac{l(S^V) \cdot s}{1-\lambda}$ sites used thus far. Since $l(\overline{v}) \leq \lambda$, this means that the SS "length" $l(B_i^V)$ of all these bins B_i will be $l(B_i^V) > 1 - \lambda$. Summing over all bins this gives:

$$\sum_i l(B_i^V) > (1-\lambda) \cdot \frac{l(S^V) \cdot s}{1-\lambda} = l(S^V) \cdot s$$

Since $\sum_i l(B_i^V) \leq s \cdot l(S^V - \{\overline{v}\})$ (by Lemma A.1), the above inequality gives: $l(S^V - \{\overline{v}\}) > l(S^V)$, which is impossible. This completes the proof of the stated upper bound.

This result basically states that, given an arbitrary collection S^V of λ-granular s-dimensional vectors and an arbitrary rule for partitioning these vectors into s-dimensional bins of unit capacity such that each

bin is filled to a height of at least $1-\lambda$, then the number of bins produced cannot exceed $\frac{l(S^V)\cdot s}{1-\lambda}$. To show that this upper bound is tight, we demonstrate a *worst-case collection* of s-dimensional demand vectors S^V and a *worst-case partitioning* of these vectors for which the stated upper bound on the required number of sites is actually reached. Consider a collection S^V of $P\cdot(k-1)$ s-dimensional vectors, where P and k are positive integers. Suppose that each vector has a non-zero value of $\frac{1}{k-\epsilon}$ ($\epsilon > 0$ a small positive number) in only one of its components and a value of zero in all other components. (Note that, all these vectors are λ-granular with $\lambda = 1/(k-\epsilon)$.) Furthermore, for every vector component i, there is an equal number (i.e., $\frac{P}{s}\cdot(k-1)$) of vectors with a non-zero value at component i, for $i = 1,\ldots,s$. Without loss of generality, we assume that P/s is an integer. Now, consider the (worst-case) partitioning of these vectors that places only vectors with the *same non-zero dimension* in the same bin. (Note that such an "inefficient" packing rule could be imposed by other scheduling criteria, for example, balancing work requirements across bins.) It is fairly easy to see that such a packing will require exactly P bins, since each bin can accommodate at most $k-1$ vectors with the same non-zero component of $1/(k-\epsilon)$. We will now show that this essentially coincides with the upper bound stated in this lemma. Our upper bound for this case is:

$$\frac{l(S^V)\cdot s}{1-\lambda} = \frac{\frac{P}{s}\cdot(k-1)\cdot\frac{1}{k-\epsilon}\cdot s}{1-\frac{1}{k-\epsilon}} = P\cdot\frac{k-1}{k-1-\epsilon},$$

which obviously converges to P as $\epsilon \to 0$. ∎

Proof of Theorem 4.2: First, note that by Lemma 4.2 PipeSched will be able to pack C in P_C. We now prove the following proposition that holds for any packing of S^W into P_C d-dimensional bins.

Proposition A.1 There exists an index $j \in \{1,\ldots,P_C\}$ such that

$$\sum_{\overline{w}\in B_j^W} l(\overline{w}) \le d\cdot\frac{l(S^W)}{P_C}$$

Proof: Assume that, to the contrary, we have:

$$\sum_{\overline{w}\in B_j^W} l(\overline{w}) > d\cdot\frac{l(S^W)}{P_C}, \text{ for all } j = 1,\ldots,P_C.$$

Summing over all j, this gives

$$\sum_{\overline{w}\in S^W} l(\overline{w}) > d\, l(S^W)$$

which contradicts Lemma A.1. ∎

Consider the work vector packing produced by PipeSched, where, without loss of generality, we assume that the sites have been renumbered in non-increasing order of total work; that is, $l(B_1^W) \ge l(B_2^W) \ge \ldots \ge l(B_{P_C}^W)$ (Figure (17)).

If $l(B_1^W) \le d\cdot\frac{l(S^W)}{P_C}$, the theorem obviously holds. Assume $l(B_1^W) > d\cdot\frac{l(S^W)}{P_C}$ Let $\overline{w}_{i_0}$ be the first work vector to "push" the total work of site B_1 over $d\cdot\frac{l(S^W)}{P_C}$. Also let $W_e =< \overline{w}_{i_0},\ldots,\overline{w}_{i_M} >$, $m \ge 0$, be the time-ordered list of vectors packed at site B_1 after that moment (including $\overline{w}_{i_0}$).

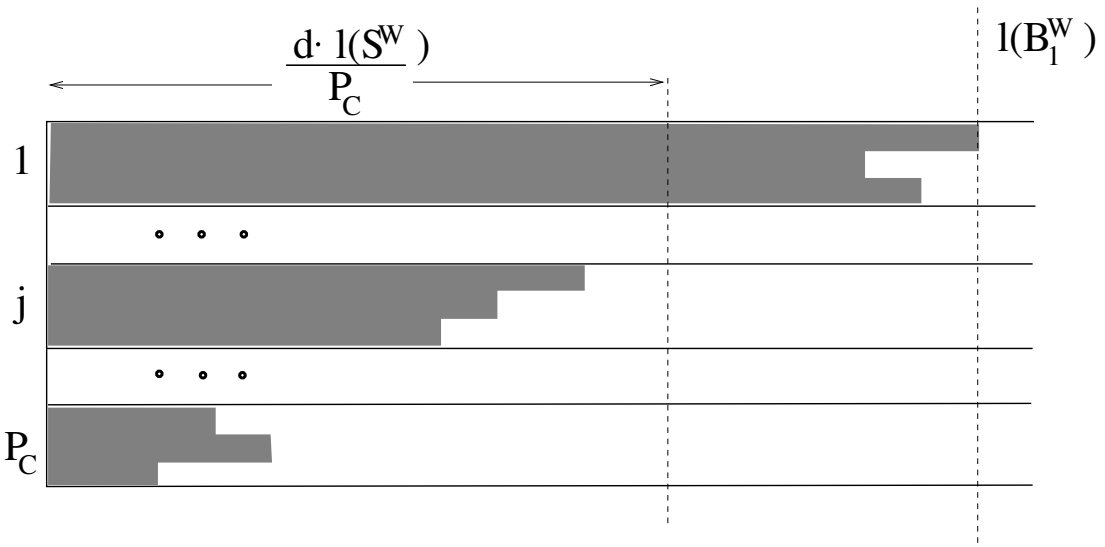

Figure 17: A packing of S's work vectors in P sites.

By Proposition A.1, we know that in the packing produced by PipeSched there exists a site B_j such that $\sum_{\overline{w}\in B_j^W} l(\overline{w}) \leq d \cdot \frac{l(S^W)}{P_C}$. By the logic of PipeSched we know that site B_j was not allowed to pack any of the vectors $\overline{w}_{i_k}$, $k = 1, \ldots, M$ only because of ss resource constraints. This implies that when the packing of $\overline{w}_{i_1}$ was taking place we had:

$$\sum_{\overline{v}\in B_j^V} l(\overline{v}) \geq l(B_j^V) > 1 - \lambda.$$

Also, by the order of packing we know that at the time of packing $\overline{w}_{i_1}$ we had:

$$\frac{l(\overline{w}_{i_1})}{l(\overline{v}_{i_1})} \leq \frac{\sum_{\overline{w}\in B_j^V} l(\overline{w})}{\sum_{\overline{v}\in B_j^V} l(\overline{v})},$$

which by the above inequality and Proposition A.1 gives:

$$\frac{l(\overline{w}_{i_1})}{l(\overline{v}_{i_1})} \leq \frac{d \cdot l(S^W)}{P_C \cdot (1-\lambda)}.$$

Again, by the order of packing we know that $\frac{l(\overline{w}_{i_1})}{l(\overline{v}_{i_1})} \geq \frac{l(\overline{w}_{i_2})}{l(\overline{v}_{i_2})} \geq \ldots \geq \frac{l(\overline{w}_{i_M})}{l(\overline{v}_{i_M})}$, which in turn implies that:

$$\frac{l(\overline{w}_{i_1})}{l(\overline{v}_{i_1})} \geq \frac{\sum_{k=1}^M l(\overline{w}_{i_k})}{\sum_{k=1}^M l(\overline{v}_{i_k})}.$$

Combining the last two inequalities and using the fact that $\sum_{k=1}^M l(\overline{v}_{i_k}) \leq s \cdot l(\sum_{k=1}^M \overline{v}_{i_k}) \leq s$ (since all these ss vectors "fit" in one site), we have:

$$\frac{\sum_{k=1}^M l(\overline{w}_{i_k})}{s} \leq \frac{d \cdot l(S^W)}{P_C \cdot (1-\lambda)},$$

or,

$$\sum_{k=1}^{M} l(\overline{w}_{i_k}) \leq \frac{d \cdot s}{1-\lambda} \cdot \frac{l(S^W)}{P_C}.$$

Thus, the response time of the schedule produced by PIPESCHED can be bounded as follows (see Figure 17):

$$T_H \leq d \cdot \frac{l(S^W)}{P_C} + \sum_{k=1}^{M} l(\overline{w}_{i_k}) + l(\overline{w}_{i_0}) \leq d \cdot (1 + \frac{s}{1-\lambda}) \cdot \frac{l(S^W)}{P_C} + T^{max}.$$

This completes the proof. ∎

Proof of Theorem 4.3: Let $S_j^W = \cup_{C \in L_j} S_C^W$ for all $j = 1, \ldots, k$, with S_j^V, S_j^{TV} defined similarly. Also, define $T_j^{max} = \max_{C \in L_j} T_C^{max}$. Finally, let H_j denote the parallel execution time of the j^{th} layer (i.e., the clones in L_j) as determined by PIPESCHED. From Theorem 4.2 we know that $H_j \leq d(1+\frac{s}{1-\lambda}) \cdot \frac{l(S_j^W)}{P} + T_j^{max}$, for all $j = 1, \ldots, k$. Thus, for the overall execution time we have:

$$T_H = \sum_{j=1}^{k} H_j \;\leq\; d(1 + \frac{s}{1-\lambda}) \cdot \sum_{j=1}^{k} \frac{l(S_j^W)}{P} \;+\; \sum_{j=1}^{k} T_j^{max}. \tag{3}$$

By the ordering of the pipelines in L we know that the total volume packed in layer j is $l(S_j^{TV}) \geq T_{j+1}^{max} \cdot l(S_j^V)$ for all $j = 1, \ldots, k-1$. Furthermore, by the condition used in the layering of the pipes (Step 2 of LEVELSCHED), we have:

$$l(S_j^{TV}) + l(S_{j+1}^{TV}) \;\geq\; T_{j+1}^{max} \cdot [l(S_j^V) + l(S_{C_{i_{j+1}}}^V)] \;>\; T_{j+1}^{max} \cdot \frac{P(1-\lambda)}{s}$$

for all $j = 1, \ldots, k-1$. Summing over all j this gives

$$2 \cdot \sum_{1}^{k} l(S_j^{TV}) \;>\; \sum_{1}^{k-1} [l(S_j^{TV}) + l(S_{j+1}^{TV})] \;>\; \frac{P(1-\lambda)}{s} \sum_{2}^{k} T_j^{max}.$$

Combining this with Inequality 3 we get:

$$T_H \;<\; d(1 + \frac{s}{1-\lambda}) \cdot \sum_{j=1}^{k} \frac{l(S_j^W)}{P} \;+\; \frac{2s}{1-\lambda} \cdot \sum_{j=1}^{k} \frac{l(S_j^{TV})}{P} \;+\; T_1^{max}.$$

Note that by the ordering of L, $T_1^{max} = T^{max}$ (the overall maximum). Using the fundamental propery of $l()$ for the two summations in the above inequality we have:

$$T_H \;<\; d^2(1 + \frac{s}{1-\lambda}) \cdot \frac{l(S^W)}{P} \;+\; \frac{2s^2}{1-\lambda} \cdot \frac{l(S^{TV})}{P} \;+\; T^{max}.$$

∎