Cost-based Partitioning for Distributed and Parallel Simulation of Decomposable Multiscale Constructive Models

Sunwoo Park, C. Anthony Hunt and Bernard P. Zeigler

SIMULATION 2006; 82; 809
DOI: 10.1177/0037549706075479

Cost-based Partitioning for Distributed and Parallel Simulation of Decomposable Multiscale Constructive Models

Sunwoo Park
BioSystems Group, Department of Biopharmaceutical Sciences University of California, San Francisco 513 Parnassus Ave, San Francisco CA 94143-0446, USA

C. Anthony Hunt
BioSystems Group, Department of Biopharmaceutical Sciences, and Joint Graduate Group in Bioengineering University of California, Berkeley and San Francisco 513 Parnassus Ave, San Francisco CA 94143-0446, USA
a.hunt@ucsf.edu

Bernard P. Zeigler
Department of Electrical and Computer Engineering University of Arizona, 1230 E. Speedway Blvd Tucson, AZ 85721, USA

We present a concise, generic, and configurable partitioning approach for decomposable, modular, and multiscale (or hierarchical) constructive models. A generic model partitioning (GMP) algorithm decomposes a given multiscale model to a set of partition blocks based on a cost modeling and analysis method in polynomial time. It minimizes model decompositions and constructs monotonically improved partitioning outcomes during the partitioning process. The cost modeling and analysis method enables translating subjective, domain-specific, and heterogeneous resource information to objective, domain-independent, and homogeneous cost information. By translating models to a homogeneous cost space and describing partitioning logics over the space, the proposed algorithm utilizes domain-specific knowledge to produce the best partitioning results without any modification of its programming logics. As a consequence of its clean separation between domain-specific partitioning requirements and goals, and generic partitioning logic, the proposed algorithm can be applied to a variety of partitioning problems in large-scale systems biology research utilizing distributed and parallel simulation. It is expected that the algorithm improves overall performance and efficiency of in silico experimentation of complex multiscale biological system models.

Keywords: Multiscale partitioning, model decomposition, discrete event systems specification (DEVS), systems biology, resource allocation

1. Introduction

Given the challenges faced by the emerging field of systems biology [1–3], multiscale constructive simulation modeling is an attractive approach for describing large, complex, multiscale biological systems. It is expected to enable representing aspects of structural and behavioral characteristics of multiscale system hierarchies of components interacting with each other and their environment. Heterogeneous and multifaceted system features can also be represented within such models. Such aggregations are often infeasible or difficult for the more traditional equation-based inductive models.

However, efficient execution of complex multiscale simulation models is challenging. The models are easily exposed to low degrees of parallelism and are also prone to unsatisfactory resource distribution to a set of computational entities (e.g. processors) in distributed and parallel computing environments. In order to increase the degree of parallelism while optimizing resource allocation and managing core modeling and simulation (M&S) issues,

SIMULATION, Vol. 82, Issue 12, December 2006 809–826

DOI: 10.1177/0037549706075479
Figure 4 appear in color online: http://sim.sagepub.com

we need to consider computational and resource management issues. Prominent among these issues are model partitioning, model deployment, remote activation, parameter sweeping and optimization, and experimentation automation [4, 5]. Model partitioning constructs a set of fine-grain component models from a coarse-grain multiscale model. Model deployment dispatches the decomposed models to the set of computational entities based on a certain heuristic. Remote activation reactively launches a simulator with a model and builds communication channels with other simulators when the model is available within the simulator's computational boundary. Parameter sweeping and optimization minimize exploration of uninterested parameter spaces. Experimentation automation pipelines a series of distinctive experimental phases to an automated workflow for the efficient execution of large-scale in silico experiments or large numbers of distinctive but repetitive experiments. Among these issues, this paper focuses on the issue of partitioning.

Multiscale model partitioning plays a key role in efficient execution of complex multiscale simulation models. By decomposing a complex multiscale model to a set of component models, it enables building and maintaining optimal model distribution over computational entities and enhances the degree of parallelism. Design and implementation of generic partitioning algorithms that can be applied to a variety of multiscale models is challenging. We must simultaneously consider two design aspects: specialization and generalization. It is desirable to use domain-specific or domain-aware knowledge to produce optimal partitioning results. However, existing partitioning algorithms use domain-independent or domain-neutral low-level information, such as execution time, communication time, delay, and memory requirements. In doing so, it is preferable to maintain generic partitioning logics that can be widely applied. However, these considerations can conflict. We adopt a cost modeling and analysis method for our partitioning algorithm in order to reduce conflicts. The method enables translating domain-specific and heterogeneous resource information into objective, domain-independent, and homogeneous cost information. The method's use leads to a class of algorithms that efficiently partition a set of decomposable multiscale models while preserving important aspects of both the specialization and generalization paradigms. The concept and related issues of this method are addressed in Section 3.

We propose a generic model partitioning (GMP) algorithm that uses the cost modeling and analysis method to decompose a modular, multiscale constructive model into a set of partition blocks in polynomial time. The algorithm describes a partitioning programming logic over a domain-independent cost space that is constructed by applying selected cost modeling and analysis techniques. The process allows the GMP algorithm to be concise, generic, and configurable. The algorithm produces high-quality partitioning outcomes with the minimum model decomposition. The quality of partitioning (QoP) is progressively improved until the best partitioning result is attained. Furthermore, it enables implementing various partitioning strategies by using different cost measures and functions instead of modifying the partitioning logics. The algorithm is described in detail in Section 4. Complexity and execution time analysis of the algorithm are presented in Section 5.

To present the usability and power of the GMP algorithm, we apply it to multiscale, decomposable, modular discrete event systems specification (DEVS) models. DEVS is a discrete-event oriented multiscale, constructive M&S approach [6–8]. It provides a solid foundation for theoretical or practical M&S driven systems biology and has been applied to multiscale biological problems [9–12]. A set of GMP DEVS partitioners has been successfully developed for large-scale distributed simulation systems [4, 13, 14]. A collection of qualitative and quantitative experimental results and their analysis are presented in Section 6.

2. Background and Related Work

Model partitioning is the process of aggregating or dividing (decomposable) models into a set of partition blocks. In distributed and parallel simulation systems, it plays vital roles in three processes: resource allocation and management, load sharing and balancing, and optimization. Performance, efficiency, and utilization can be significantly improved by optimally distributing models into active or passive system entities (e.g. simulators and coordinators). Optimal distribution is closely related to how models are partitioned and deployed. Thus, it is important to develop algorithms that produce optimal or, at least, acceptable partitioning results with respect to the end points of interest. However, most model partitioning algorithms focus on non-decomposable models that are formulated as a graph or a hyper-graph structure [15–18]. Multilevel partitioning algorithms transform the structure into a hierarchical alternative [19–23]. Neither deals with decomposable models. We can produce better partitioning results [4, 24, 25]. As model complexity increases, they have naturally evolved into hierarchical and modular structures. Such evolution escalates the demand for new classes of partitioning algorithms that efficiently handle those structures.

Partitioning algorithms are mainly divided into three main classes: random partitioning, partitioning refinement, and heuristic. Random partitioning algorithms randomly aggregate or segregate models to a set of partition blocks. Partitioning refinement algorithms improve partitioning results during the partitioning process. Heuristic partitioning algorithms utilize domain-specific knowledge or particular optimization techniques.

The Kernighan–Lin (KL) algorithm is an example of random partitioning combined with partitioning

refinement. The KL algorithm initially builds a partitioning result by randomly assigning models to partition blocks; it then revises the quality of the results by swapping models between those blocks whenever swapping produces a better partitioning result [26]. The performance of the KL algorithm has been improved from $O(n^3)$ to $O(\max\{E \cdot \log n, E \cdot \deg_{\max}\})$ by Dutt [27], and to $O(E)$ by Fiduccia and Mattheyeses [28], and to $O(V + E)$ by Diekmann, Monien, and Preis [29]. V, E, and $\deg_{\max}$ are the total number of vertices, the total number of edges, and the maximum node degree, respectively. The quality of partitioning is substantially bound to the initial partitioning result. Thus, it is desirable to incorporate domain specific heuristics to improve result quality [30].

Multifarious heuristics have been applied to model partitioning algorithms. Structural and spatial relationships between models are used in recursive bisection algorithms. The algorithms split a graph into two subgraphs and recursively bisect each subgraph based on particular geometric information. Recursive coordinate bisection (RCB), recursive inertial bisection (RIB), and orthogonal recursive bisection (ORB) algorithms use the property of spatial orthogonality: a coordinate axis, an axis of angular momentum, and an orthogonal plane to the axis [31–33]. Recursive graph bisection (RGB) algorithms use the shortest path length between two graph nodes [34]. Recursive spectral bisection (RSB) and eigenvector recursive bisection (ERB) algorithms use an eigenvector representing connectivity and distance between nodes [36–39]. Various optimization techniques including simulated annealing (SA), mean field annealing (MFA), Tabu search (TS), and genetic algorithm (GA) have been also applied to model partitioning algorithms [40–44].

Hierarchical partitioning works by either decomposing or building hierarchical structures based on specified decision-making criteria. Hierarchical structure is commonly represented by a multilevel, acyclic, directed graph (ADG) or a tree structure. During the partitioning process, the hierarchical structure is dynamically created and updated over time and space. A partitioning policy specifies how and when the structure is updated. Three widely used policies are flattening, deepening, and heuristic. Flattening is a structural decomposition technique that transforms the hierarchical structure into a non-hierarchical structure. Deepening, also known as hierarchical clustering, is a structural aggregation technique that transforms a non-hierarchical structure into a hierarchical one. Heuristic is any technique other than flattening and deepening. In this paper, we refer to partitioning algorithms based on the flattening and deepening approaches as multiscale and multilevel partitioning algorithms, respectively. Multilevel partitioning has been investigated extensively over the past few decades [19–23]. However, multiscale partitioning has received less attention.

In this paper, we reduce the scope of multiscale partitioning algorithms to random, ratio-cut, and heuristic. For a given hierarchical and decomposable cost tree that preserves the structural relationship between the components of a DEVS coupled model (as shown in Figure 1), a random algorithm decomposes the tree and randomly assigns nodes or subtrees to a set of partition blocks. A ratio-cut algorithm cuts a subtree that has the minimum cost disparity compared to the average cost of the tree. The average cost of the tree is computed by dividing the cost of the root node of the tree by the requested number of partition blocks. Once a subtree is assigned to a partition block, the average cost is recomputed while excluding the subtree. This is repeated until only one partition block is left. The last partition block is populated with the remaining nodes that are not assigned to other partition blocks. The HIPART algorithm is an example of the ratio-cut algorithm [24]. A heuristic algorithm is one that uses any technique other than random and ratio-cut approaches. The ENCLOSURE algorithm is an example of the heuristic algorithm [25].

3. Cost Modeling and Analysis Method

The cost modeling and analysis method provides a means of transforming heterogeneous resource information into homogeneous cost information while conducting analyses over a cost space. A "cost" is a homogeneous object representing heterogeneous resource information (e.g. single value, a set of discrete objects, and a continuous range). A cost "measure" is a conceptual metric that captures heterogeneous resource information in terms of cost (e.g. complexity, I/O connectivity, dynamic activity, and latency). Because a cost measure is a parametric method subject to certain axioms, algorithms based on the method are generic and applicable to any family of computational tasks (e.g. constructive simulation models) provided that there is a way to manipulate the appropriate cost information. However, a more general concept potentially includes other important determiners of a task such as number of messages sent and received. By applying one or more cost measures, a task is abstracted to a cost regardless of its complexity or heterogeneity. The homogeneity of the cost allows the proposed algorithm to be applicable to heterogeneous problems by simply switching cost measures, without any modification of the algorithm itself (see Table 1). This is because of the homogeneous nature of the method. Thus, the proposed algorithm is highly adaptable and can be applied within various application domains. A cost function is a mathematical function that quantifies or qualifies resource information to cost based on a set of cost measures. Some of the operations considered in cost modeling and analysis are cost extraction, cost generation, cost aggregation, cost evaluation, and cost analysis [4].

A "cost tree" is a homomorphic representation of a decomposable, modular, and multiscale task from the perspective of cost modeling and analysis. A node in the tree is classified as either atomic or coupled. An atomic node

Table 1. An example of cost measures and cost functions

Cost measure	Cost function	Decision-making criteria
I/O connectivity	$\|X_{model}\| * \|Y_{model}\|$	The cost of a system is generally proportional to the number of I/O interfaces if the system is dedicated to serving I/O requests.
System complexity	$\|\Gamma_{model}\|$	The cost of a system is represented by the number of internal states rather than the number of I/O access points if system performance relies on its complexity.
I/O and system complexity	$\|X_{model}\| * \|Y_{model}\| * \|\Gamma_{model}\|$	The cost of a system can be captured more appropriately by considering both I/O interfaces and system complexity.
System activity	$\|\Delta_{model}\|$	The cost of a system can be captured more appropriately by considering dynamic system behaviors.

X_{model} is a set of input interfaces, Y_{model} is a set of output interfaces, Γ_{model} is a set of internal states, Δ_{model} is a set of internal transitions for a certain period of time, and $|E|$ is a counting operator that returns the total number of elements in the given set E

is a terminal node containing no child nodes. A coupled node is a non-terminal node holding at least one child node. A set of decomposable multiscale tasks are easily translated to a cost tree by applying cost functions with appropriated cost measures. Each node contains a cost (or a task and cost pair). The cost of an atomic node is generally equal to the cost of the model with which it is associated. However, the cost of a coupled node is the aggregated cost of that node and all descendants that can be reached through the tree hierarchy. Thus, the cost of a subtree starting from a particular node is acquired by simply retrieving the cost of the node without further expansion or exploration of the tree. So doing considerably reduces the amount of time and space required for parsing all descendants of the node and aggregating their costs during the cost evaluation process. We show the GMP algorithm based on the method runs in polynomial time in Section 5.

Let D be a finite discrete set of models. D refers to a decomposable set if it contains at least one model that can be expanded into a set of submodels. If D is populated with more than one model, $\{d_1, \ldots, d_n\}$, we repopulate D with a new virtual coupled model d_0 that contains all existing models in D. That is, $D = \{d_0\}$ where $d_0 = \{d_1, \ldots, d_n\}$. A cost tree T is a tree structure that represents each model $d_i \in D$ by a cost node $a_i \in A$ while preserving structural properties as shown in Figure 1. A is a set of cost nodes representing T. Every d_i is translated to a_i by a cost evaluation function, $f_{eval} : D \rightarrow A$. If d_i is a decomposable model, it is also legitimate to alternatively use a cost aggregation function, $f_{aggr} : E \rightarrow A$. E is a subset of D that contains only decomposable models in D. f_{aggr} computes the cost of a coupled component $e_i \in E$ by aggregating all costs of its children. We can build distinct collections of cost trees by applying different aggregation methods (e.g. summation, max, and average) to d_0. So doing enables delineating a multiscale model from various different perspectives based on aggregated cost. During the tree construction process, D shrinks when a decomposable model is removed and grows when the removed model is expanded and its components are added back to D.

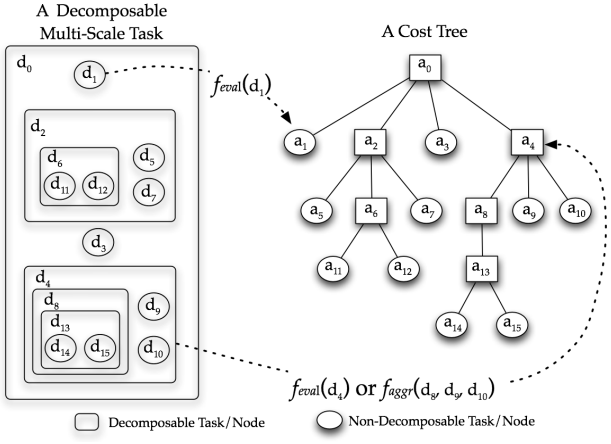


Figure 1. Cost tree construction with a cost evaluation function and/or a cost aggregation function: A cost tree T is constructed from a decomposable task D. The cost $a_i \in T$ is computed by a cost evaluation function $f_{eval} : D \rightarrow A$. If d_i is a decomposable model, the cost a_i can also be computed by a cost aggregation function $f_{aggr} : E \rightarrow A$, instead of f_{eval}. E is a subset of D that contains only decomposable models in D.

The GMP algorithm requires a cost tree. In general, the cost tree preserves the structural relationship between components of a reference model. However, variant cost trees can be constructed from the model by excluding some components of the model and also distorting structural relationship between components of the model. It enables building a set of different cost tree topologies from the model. For a given cost tree, the algorithm tries to produce the best partitioning result without any involvement in cost tree construction and validation. Difficulty, ambiguity, and accuracy of obtaining, generating, and aggregating cost information in cost tree construction are isolated into cost modeling issues. By excluding those issues from partitioning algorithmic logics, the algorithm remains concise but generic for a wide range of applications.

4. Generic Model Partitioning

A GMP algorithm decomposes a given multiscale model (e.g. a coupled model in DEVS) into a set of partition blocks. It decomposes a set of models only if model decomposition produces a better partitioning result. With the minimization of model decomposition, the GMP algorithm becomes less sensitive to the depth or the complexity of the models. Minimization makes the algorithm more flexible and scalable than other partitioning algorithms based on full decomposition. A unique feature of the GMP algorithm is its support of incremental QoP improvement during the partitioning process. This property guarantees that partitioning outcomes will evolve into better alternatives without any degradation of QoP until a best partitioning result is attained. Incremental improvement enables the GMP algorithm to produce a high degree of QoP for the given model. The GMP algorithm divides into two subalgorithms: initial partitioning and evaluation–expansion–selection (E^2S) partitioning.

4.1 Initial Partitioning

The initial partitioning algorithm constructs P partition blocks from a cost tree T. Each partition block contains at least one node. The algorithm consists of four phases: initialization, expansion, fill, and distribution, as shown in Algorithm 1. All necessary data structures are created with initial values in the initialization phase (lines 3–4). $clist$ is the list containing cost nodes. It grows and shrinks, respectively, when a node is expanded from T and is assigned to a partition block. Initially, $clist$ is populated with child nodes of a root node and every partition block is empty. If $P > |clist|$, at least one node expansion occurs until $|clist|$ becomes equal to or larger than P (lines 6–12). Node expansion is a sequence of (i) identifying and removing $n_{highest}$, which is the node having the highest cost in $clist$, (ii) expanding it, and (iii) restoring its child nodes back to $clist$. If $P \leq |clist|$, select $n_{highest}$ and assign it to an empty partition block until there exist no empty partition blocks (lines 14–16). Finally, remaining nodes in $clist$ are distributed to non-empty partition blocks until $|clist|$ becomes zero (lines 18–20). The node having the lowest cost in $clist$, n_{lowest}, is used instead of $n_{highest}$ in the distribution phase. Initial partitioning minimizes cost disparity between partition blocks by assigning a node to each empty block in a descending order and distributing remaining nodes to partition blocks in an ascending order. Initial partitioning results of the cost tree in Figure 1 over various P are provided in Figure 2.

4.2 Evaluation–Expansion–Selection Partitioning

Evaluation–expansion–selection (E^2S) partitioning improves the quality of partition results until no better result is attained. The algorithm consists of six phases: initialization, identification, expansion, fill, distribution, and evaluation, as shown in Algorithm 2. All necessary data structures are created with initial values in the initialization phase (lines 3–5). $parray$ and $earray$ are, respectively, a set of partition blocks that contain previous and next partitioning outcomes. $epartition$ is the partition block that contains the highest cost in $earray$, $PB_{highest}$. $enode$ is the coupled node that has the highest cost over all other nodes in $PB_{highest}$. The initial partitioning result is assigned to $parray$. After initialization, $enode$ is identified from $PB_{highest}$ (lines 7–18). The selected $enode$ is expanded and $epartition$ is filled with a node if it has only $enode$ (line 20 and lines 22–24). The remaining nodes

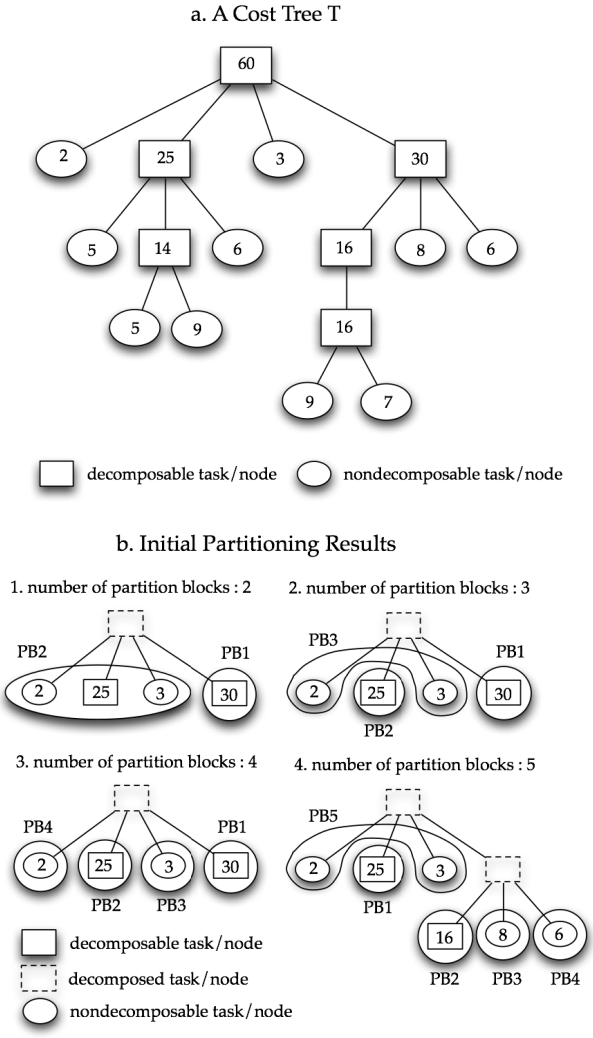


Figure 2. Initial partitioning results of the cost tree T over various P: $f_{eval}(d_i)$ = system activity of d_i, $f_{aggr}(d_i) = \sum_j f_{eval}(d_{ij})$, $d_{ij} \in d_i$, and $2 \leq P \leq 5$.

Algorithm 1. Initial partitioning algorithm ($\mathbb{A}_{Init}$)

Input:
T: a cost tree, P: a number of partition blocks
Return:
$parray$: partitioning result
Acronym:
PB: a partition block, PB_{empty}: an empty PB, $|PB| = 0$
PB_{lowest}: a PB having the lowest cost, $PB_{highest}$: a PB having the highest cost
$Node_{lowest}$: a node having the lowest cost, $Node_{highest}$: a node having the highest cost
$Node_{coupled}$: a coupled node, $Node_{coupled}^{highest}$: a coupled node having the highest cost
Operators:
$removeFrom(node, clist)$: remove a $node$ from $clist$; $node \leftarrow removeFrom(node, clist)$
$addTo(node, PB)$: add a $node$ to a partition block, PB; $PB' \leftarrow addTo(node, PB)$
$expand(node)$: expand a $node$; a set of child nodes of the $node \leftarrow expand(node)$

```
1   procedure PB[] INITIAL-PARTITIONING (CostTree T, int P)
2:    // PHASE 1: initialize clist and parray
3:    clist := child nodes of a root node in T                          ▷|clist| > 0
4:    parray := PB[P] // create P empty partition blocks                ▷∀i, |parray[i]| = 0, 1 ≤ i ≤ P
5:    // PHASE 2: expand node(s), if necessary
6:    while lengthOf(clist) < numberOf(parray) do
7:      if clist contains at least one Node_coupled
8:        clist := clist + expand(removeFrom(Node_coupled^highest, clist))
9:      else
10:       return error("cannot expand ...")                             ▷ (∄ Node_coupled ∈ clist) ∨ |clist| = 0
11:     end if
12:   end while                                                         ▷|clist| ≥ |parray|
13:   // PHASE 3: fill empty partition blocks
14:   while parray contains an PB_empty do
15:     addTo(removeFrom(Node_highest, clist), PB_empty)
16:   end while                                                         ▷∀i, |parray[i]| > 0
17:   // PHASE 4: distribute nodes in clist into partition blocks
18:   while clist is not empty do
19:     addTo(removeFrom(Node_lowest, clist), PB_lowest)
20:   end while                                                         ▷|clist| = 0
21:   return parray
22: end procedure
```

in $clist$ are distributed to non-empty partition blocks until $|clist|$ becomes zero (lines 26–28). Finally, E^{2}S partitioning recursively performs until a best result is attained (lines 30–34). If the new partitioning result $earray$ is superior to the previous result $parray$, E^{2}S partitioning continues. Otherwise, it returns $parray$ as the best partitioning result. $superiorTo()$ is a user-provided function that compares $earray$ to $parray$. An example of E^{2}S partitioning results is presented in Figure 3. Figure 3(b) illustrates monotonic QoP improvement of E^{2}S partitioning result.

5. Algorithm Analysis

To simplify our analysis, we assume a coupled model D that is translated to a cost tree $T(d, k, n)$. d is the depth of $T(d, k, n)$, k is the number of child nodes per coupled node, and n is the total number of atomic nodes. n is k^i, where $i \in 1, \ldots, d$. The total number of nodes in $T(d, k, n)$ ranges from $\sum_{i=0}^{d-1} k^i + k$ and $\sum_{i=0}^{d} k^i$ since there exists $\sum_{i=0}^{d-1} k^i$ coupled nodes, where $d > 1$ and $k > 1$. In this paper, we do not consider the complexity of constructing $T(d, k, n)$ because the construction process is not part of the GMP algorithm.

5.1 Initial Partitioning Algorithm

The length of the $clist$ after i node expansions l_i is

$$l_i = \begin{cases} \xi_0, & i = 0 \\ l_{i-1} + \xi_i, & i \geq 1, \end{cases} \quad (1)$$

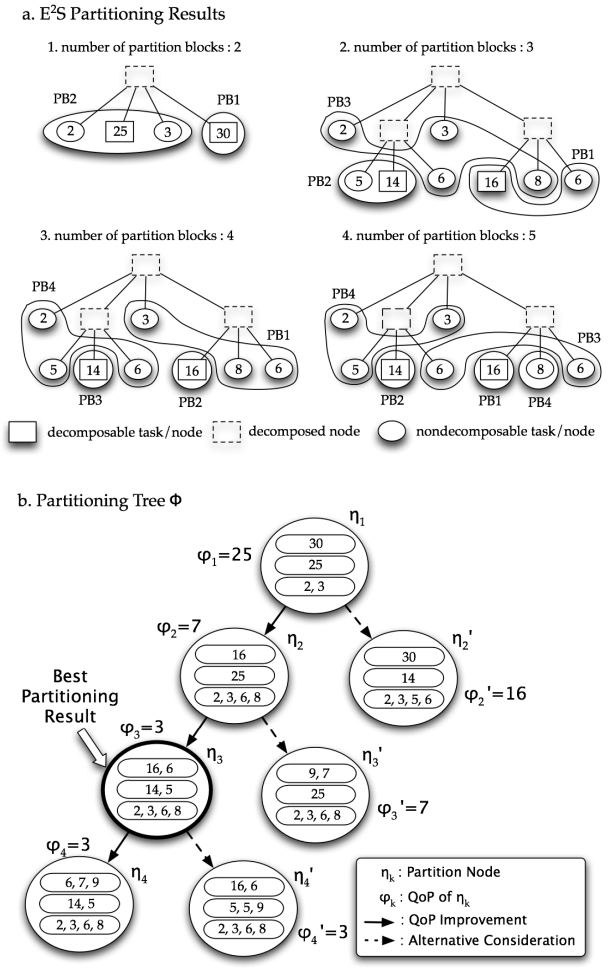


Figure 3. E²S partitioning results for various P and an example of a partitioning tree. The top figure presents partitioning results when P ranges from 2 to 5. The bottom figure presents a partitioning tree Φ when P is 3 and cost disparity between PB_{max} and PB_{min} is used as a partitioning quality measure, φ. $\varphi = |f_{aggr}(PB_{max}) - f_{aggr}(PB_{min})|$. $f_{aggr}(PB_i) = \sum_j a_j$, $1 \le j \le |PB_i|$ and $a_j \in PB_i$. PB_{max} and PB_{min} are PB_i that respectively satisfy $f_{aggr}(PB_i) \ge f_{aggr}(PB_j)$ and $f_{aggr}(PB_i) \le f_{aggr}(PB_j), \exists i \forall j((i \ne j) \wedge (1 \le i, j \le P))$. Lower φ means better QoP.

where ξ_0 is the number of children of a root node in T and ξ_i is the number of children of the expanded node after i expansions. By substituting l_{i-1} by the sum of ξ up to $(i-1)$th expansion, we can rewrite l_i as

$$\begin{aligned} l_i &= l_{i-1} + \xi_i = \dots \\ &= \sum_{j=0}^{i-1} (\xi_j - 1) - 1 + \xi_i \\ &= \sum_{j=0}^{i} \xi_j - i, \qquad i \ge 1. \end{aligned} \tag{2}$$

Assume that node expansions occur E times to guarantee $|clist| \ge |parray|$. Then, by applying equation (2) to the *while* loop (line 6, Algorithm 1), we can rearrange the conditional part of the loop to $(\sum_{j=0}^{E} \xi_j - E) < P$ (see Appendix A). E is the total number of expansions and P is the the number of partition blocks, $|parray|$.

To make analysis simple, let ξ_i be a constant value K (i.e. K-ary tree). For a given $K \in \mathbb{N}$, the conditional part is simplified to $E < (P-K)/(K-1)$ by substituting K for ξ_i (see Appendix B). Because $E \in \mathbb{N}$, the total number of node expansions needed in the initial partitioning becomes $\lceil (P-K)/(K-1) \rceil$, $1 < K < P$. No expansion occurs when $K \ge$ P. The length of $clist$ after the expanding phase, l_E, is described by P and K by substituting $\lceil (P-K)/(K-1) \rceil$ for i and K for ξ_i in equation (2). l_E is K when no expansion occurs because the $clist$ is initially populated with the children of the root node.

$$l_E = \begin{cases} (K-1) \cdot \left\lceil \frac{P-K}{K-1} \right\rceil + K, & 1 < K < P \\ K, & K \ge P. \end{cases} \tag{3}$$

P comparisons occur in the filling phase because every empty partition in the $parray$ is filled with a cost node that is extracted from the $clist$. $(l_E - P)$ comparisons occur in the distribution phase because the remaining $clist$ nodes are distributed into the $parray$.

Definition 1 For execution time complexity analysis of $\mathbb{A}_{initial}$, we define

$\xi(n)$: expand the node, n;
δ_{part}^{init}: time required for executing $\mathbb{A}_{initial}$;
$\delta(clist, nodes, add)$: time required for adding nodes to the $clist$;
$\delta(clist, nodes, remove)$: time required for removing nodes from the $clist$;
$\delta(parray, size, create)$: time required for creating the parray with size empty blocks;
$\delta(PB_i, nodes, add)$: time required for adding nodes to the PB_i.

Algorithm execution time is the sum of time spent in each phase of the algorithm. That is, $\delta_{part}^{init} = \delta_{init} + \delta_{expand} + \delta_{fill} + \delta_{dist}$. We can rewrite this as

$$\delta_{part}^{init} = \begin{cases} \delta_{init} + \sum_{i=1}^{E} \delta_{expand_i} + \sum_{i=1}^{P} \delta_{fill_i} + \sum_{i=1}^{l_E - P} \delta_{dist_i}, \\ 1 < K < P \\ \delta_{init} + 1 + \sum_{i=1}^{P} \delta_{fill_i} + \sum_{i=1}^{K-P} \delta_{dist_i}, \\ K \ge P \end{cases}, \tag{4}$$

where $\delta_{init} = \delta(clist, \xi(Node_{root}), add) + \delta(parray, P, create)$, $\delta_{expand_i} = \delta(clist, Node_{coupled}^{highest}, remove) + \delta(clist, \xi(Node_{coupled}^{highest}), add)$, $\delta_{fill_i} = \delta(clist, Node_{highest}, remove) + \delta(PB_{empty}, Node_{highest}, add)$, and $\delta_{dist_i} = \delta(clist, Node_{lowest}, remove) + \delta(PB_{lowest}, Node_{lowest}, add)$. By applying E and l_E to equation (4), we obtain

Algorithm 2. E^{2}S partitioning algorithm ($\mathbb{A}_{E^2S}$)

Input:
$parray$: previous partitioning result
Return:
$parray$: new partitioning result
Operators:
$evaluate(PB)$: evaluate partition blocks; $value \leftarrow evaluate(PB)$
$superiorTo(PB_1, PB_2)$: check PB_1 is superior to PB_2;
$True$ or $False \leftarrow superiorTo(PB_1, PB_2)$

```
1   procedure EVALUATION-EXPANSION-SELECTION PARTITIONING(PB parray)
2:    // PHASE 1: initialize earray and epartition
3:    earray := parray                                   ▷∀i, |earray[i]| > 0
4:    epartition := PB_highest in earray                 ▷ epartition ≠ ∅
5:    enode := null                                      ▷ enode = ∅
6:    // PHASE 2: identify an expandable PB from earray
7:    while true do
8:      if epartition = null return parray               ▷∀i, ∄Node_coupled ∈ earray[i]
9:      else
10:       if epartition contains Node_coupled then
11:         enode := Node_coupled^highest in epartition
12:         break                                        ▷enode ≠ ∅
13:       else
14:         epartition := select the PB_highest from earray
15:           excluding previously selected PBs
16:       end if
17:     end if
18:   end while                                          ▷∃i, enode ∈ earray[i]
19:   // PHASE 3: expand enode and put them into clist
20:   clist := expand(removeFrom(enode, epartition))     ▷ |clist| = |enode|
21:   // PHASE 4: fill the epartition with Node_highest if epartition is empty
22:   if epartition is empty then
23:     addTo(removeFrom(Node_highest, clist), epartition)
24:   end if                                             ▷∀i, |earray[i]| > 0
25:   // PHASE 5: distribute nodes to earray
26:   while clist is not empty do
27:     addTo(removeFrom(Node_lowest, clist), PB_lowest)
28:   end while                                          ▷|clist| = 0
29:   // PHASE 6: evaluate a new partitioning result
30:   if superiorTo(evaluate(earray), evaluate(parray)) then
31:     return Evaluation-Expansion-Selection Partitioning(earray)   ▷φ(earray) > φ(parray)
32:   else
33:     retrun parray                                    ▷ φ(earray) ≤ φ(parray)
34:   end if
35: end procedure
```

$$\delta_{part}^{init} = \begin{cases} \delta_{init} + \sum_{i=1}^{\left\lceil \frac{P-K}{K-1} \right\rceil} \delta_{expand_i} + \sum_{i=1}^{P} \delta_{fill_i} \\ + \sum_{i=1}^{(K-1)\left\lceil \frac{P-K}{K-1} \right\rceil + K - P} \delta_{dist_i}, \\ 1 < K < P \\ \delta_{init} + 1 + \sum_{i=1}^{P} \delta_{fill_i} + \sum_{i=1}^{K-P} \delta_{dist_i}, \\ K \geq P \end{cases} . \quad (5)$$

To make analysis of the algorithm simple, assume it takes one time unit either to run an operator or to evaluate a conditional statement. Then, δ_{init} takes three units: one unit for the expansion of the root node and two units for the initialization of $clist$ and $parray$. δ_{expand_i} takes five units: two units for the evaluation of conditional parts of both $while$ and if loops and three units for the execution of the $remove\text{-}expand\text{-}add$ operation. Both δ_{fill_i}

and δ_{dist_i} take three units: one unit for the evaluation of conditional part of the *while* loop and two units for the execution of the *remove-add* operation. By substituting all $\delta(\cdot)$ in equation (5) by appropriate the execution time,

$$\delta_{part}^{init} = \begin{cases} 3 + \left\lceil \frac{P-K}{K-1} \right\rceil \cdot 5 + P \cdot 3 \\ + \left((K-1) \left\lceil \frac{P-K}{K-1} \right\rceil + K - P \right) \cdot 3, \\ 1 < K < P \\ 3 + 1 + P \cdot 3 + (K-P) \cdot 3, \\ K \geq P \end{cases} . \quad (6)$$

By rearranging equation (6), we obtain

$$\delta_{part}^{init} = \begin{cases} (3K+2) \left(\left\lceil \frac{P-K}{K-1} \right\rceil + 1 \right) + 1, & 1 < K < P \\ 3K + 4, & K \geq P \end{cases} . \quad (7)$$

Equation (7) shows that, for a given partition block size P, the total execution time of the initial partitioning algorithm is more sensitive to the number of children of a coupled node (e.g. K in K-ary tree) rather than the total number of components or the spatiotemporal complexity of each component in a given multiscale model. It also implies that the performance of the algorithm is highly bound to the number of expansions, E, rather than model complexity.

5.2 E²S Partitioning Algorithm

Definition 2 For execution time complexity analysis of $\mathbb{A}_{E^2S}$, we define

ξ_{enode}: the number of child nodes expanded from *enode*, $|\xi(enode)|$;

γ: the number of recursions until a best partitioning result is attained;

$\delta_{part}^{E^2S}$: time required for executing $\mathbb{A}_{E^2S}$.

The E²S partitioning algorithm has six phases as described in Algorithm 2. Thus, the total execution time of the algorithm is the sum of time spent in those phases: $\delta_{part}^{E^2S} = \delta_{init} + \delta_{identify} + \delta_{expand} + \delta_{fill} + \delta_{dist} + \delta_{eval}$. In the *while* loop, l comparisons occur to find the *enode* in the identification phase, $1 \leq l \leq$ P. l divides into $l-1$ for partition blocks having no coupled node and 1 for the partition block having at least one coupled node. At most, ξ_{enode} comparisons occur in the distribution phase to redistribute all nodes of the *clist* to the *earray*. No comparisons occur in other phases.

$$\begin{aligned} \delta_{part}^{E^2S} &= \delta_{init} + \sum_{i=1}^{l-1} \delta_{identify,\neg enode} \\ &+ \delta_{identify,enode} + \delta_{expand} \\ &+ \varepsilon \cdot \delta_{fill} + \sum_{i=1}^{\xi_{enode}-\varepsilon} \delta_{dist_i} + \delta_{eval}, \end{aligned} \quad (8)$$

where l is the total number of comparisons occurred in the while loop to find *enode*, ε is 1 if *epartition* is empty after the expansion phase and otherwise 0, $\delta_{identify,\neg enode}$ is the time need for handling *epartition* having no coupled node, and $\delta_{identify,enode}$ is the time needed for handling *epartition* having at least one coupled node.

To simplify the analysis, assume it takes one time unit to run an operator or to evaluate a conditional statement. Then, δ_{init} takes three units to initialize *earray*, *epartition*, and *enode*. $\delta_{identify,\neg enode}$ takes four units to handle *epartition* having no coupled node. $\delta_{identify,enode}$ takes five units to identify *enode*. δ_{expand}, δ_{fill}, and δ_{dist_i} take three units, respectively. δ_{eval} takes three + $\delta_{part}^{'E2S}$ units. $\delta_{part}^{'E^2S}$ is equal to the time needed to run the algorithm recursively until the best result is attained. $\delta_{part}^{'E^2S}$ is 1 if $evaluate(earray)$ is not better than $evaluate(parray)$. By substituting all $\delta(\cdot)$ in equation (8) by the appropriate execution time, we obtain

$$\begin{aligned} \delta_{part}^{E^2S} &= 3 + (l-1) \cdot 4 + 5 + 3 + \xi_{enode} \cdot 3 \\ +3 + \delta_{part}^{'E^2S} &= 4l + 3\xi_{enode} + 10 + \delta_{part}^{'E^2S}. \end{aligned} \quad (9)$$

Assume $\delta_{part}^{'E^2S}$ runs γ times recursively to attain a best result. Then, we rewrite equation (9) as

$$\begin{aligned} \delta_{part}^{E^2S} &= 4l + 3\xi_{enode} + 10 \\ &+ \sum_{i=1}^{\gamma} \left(4l_i + 3\xi_{enode}^i + 10 \right) + 1, \end{aligned} \quad (10)$$

where l_i is l at the ith recursion in $\delta_{part}^{'E^2S}$, and ξ_{enode}^i is ξ_{enode} at the ith recursion in $\delta_{part}^{'E^2S}$.

In most cases, l_i is 1. However, it could vary dynamically depending on the content of the $parray$. We approximate l_i by introducing $\bar{l}$, the average of l_i, as

$$\delta_{part}^{E^2S} = \sum_{i=0}^{\gamma} \left(4\bar{l} + 3\xi_{enode}^i + 10 \right) + 1, \quad (11)$$

where $\bar{l}$ is the average of l, $1/(\gamma+1) \sum_{i=0}^{\gamma} l_i$, and ξ_{enode}^i is ξ_{enode} at the ith recursion

For a K-ary cost tree, we rewrite equation (11) by substituting ξ_{enide}^i by K as follows:

$$\delta_{part}^{E^2S} = \sum_{i=0}^{\gamma} \left(4\bar{l} + 3K + 10 \right) + 1. \quad (12)$$

Equation (12) implies that the total execution time of the E²S partitioning algorithm is sensitive to the degree of QoP rather than the total number of components or the spatiotemporal complexity of each component in a given multiscale model.

5.3 Worst Case Analysis

5.3.1 Initial Partitioning Algorithm

For a worst case in the initial partitioning, assume the cost tree $T(d,k,n)$ sustains the following constraints: (i) the total number of atomic nodes n is k^d; (ii) the number of partition blocks P is k^d; (iii) $d \gg k$. E is induced to $(k^d-1)/(k-1)$ from the constraints i and ii. The constraints imply that there exist $\sum_{i=0}^{d-1} k^i$ coupled nodes and that they are all expanded. $\sum_{i=0}^{d-1} k^i$ is simplified to $(k^d-1)/(k-1)$ using the geometric series. l_E is computed to k^d+k-1 by substituting $(k^d-1)/(k-1)$ for i and k for ξ_i (see equation (2)). By substituting all $\delta(\cdot)$ in equation (4) by the appropriate execution time and rearranging it, we obtain

$$\begin{aligned}\delta_{part}^{init} &= 3k^d + 5\frac{k^d-1}{k-1} + 3k \\ &= 3n + 5\frac{n-1}{n^{1/d}-1} + 3n^{1/d}. \end{aligned} \quad (13)$$

As k is $n^{1/d}$ and $n \gg d$, $O(\delta_{part}^{init})$ is $O(n)$ for $T(d,k,n)$, $d \gg k > 1, n = k^d$.

5.3.2 E^{2}S Partitioning Algorithm

For the worst case in the E^{2}S partitioning algorithm, assume an additional constraint: (iv) a new partitioning result is always superior to the previous one.

In the case, l_i is 1 because the $PB_{highest}$ always contains at least one coupled node. We can compute γ for the given $T(d,k,n)$ by

$$\gamma = \sum_{i=1}^{d-1} k^i = \frac{k^d-1}{k-1} - 1,\ k > 1. \quad (14)$$

By substituting l_i and γ with 1 and $(k^d-1)/(k-1)-1$, respectively, in equation (12), we obtain

$$\delta_{part}^{E^2S} = \left(\frac{k^d-1}{k-1} - 1 + 1\right)\cdot(4+3k+10) + 1. \quad (15)$$

We rewrite equation (15) in terms of n:

$$\delta_{part}^{E^2S} = 3\cdot\left(\frac{n-1}{n^{1/d}-1}\right)\cdot\left(n^{1/d}+\frac{14}{3}\right) + 1. \quad (16)$$

$O(\delta_{part}^{E^2S})$ is $O(n)$ for the cost tree $T(d,k,n)$, $d \gg k > 1$, $n = k^d$.

5.4 Parameter Optimization

From an algorithm analysis perspective, if an algorithm needs a set of parameters to perform its task, it is important to know which parameter values produce the minimal algorithm execution time as well as to know the algorithm execution time for a given parameter value set. Searching optimal parameter value sets from a large parameter space produced by the parameter set can be challenging.

As part of parameter optimization analysis in the GMP algorithm, we discuss here an optimal P for a fixed T and an optimal T for a fixed P. This helps to predict or understand the relationship between P and T. For example, we can determine an optimal P for a particular multiscale model and an optimal structure of a multiscale model (e.g. K in K-ary cost tree) for a particular P. Optimal P and T do not necessarily mean that a simulation will run fast. However, they can attest it if the cost of each node in T represents the execution time of a component in a reference model.

5.4.1 Optimal Number of Partition Blocks

For a given coupled model, the optimal number of partition blocks P_{opt} is identified by searching P that produces minimum execution time from the following equation. $\delta_P^K(\mathbb{A}_{Init})$ is equation (7) that represents the execution time of the initial partitioning algorithm for a fixed K and an arbitrary P:

$$\min_{1<P\le\infty}\{\delta_P^K(\mathbb{A}_{Init})\} = \begin{cases} \min_{K<P\le\infty}\{(3K+2)\left(\left\lceil\frac{P-K}{K-1}\right\rceil+1\right)+1\}, \\ K < P \le \infty \\ 3K+4, \quad P \le K \end{cases}. \quad (17)$$

By rearranging equation (17), we obtain

$$\min_{1<P\le\infty}\{\delta_P^K(\mathbb{A}_{Init})\} = \begin{cases} \min_{K<P\le\infty}\{\frac{3K+2}{K-1}P + \frac{3K+2}{K-1}+1\}, \\ K < P \le \infty \\ 3K+4, \quad P \le K \end{cases}. \quad (18)$$

P_{opt} is $K+1$ when $K < P \le \infty$ because $\delta_P^K(\mathbb{A}_{Init})$ grows linearly as P increases. $\delta_{K+1}^K(\mathbb{A}_{Init})$ produces minimum execution time. P_{opt} is P when $P \le K$ because $\delta_P^K(\mathbb{A}_{Init})$ is governed by only K independent from P:

$$P_{opt} = \begin{cases} K+1, & K < P \le \infty \\ P, & P \le K \end{cases}. \quad (19)$$

Table 2. Cost patterns for generating various computational workload distributions of a model [45–48]

Pattern	PMF	Parameters	Distribution	Load
$C_{unitstep}$	$\delta(x)$	None	Unit step	Low
C_{exp}	$\lambda e^{-\lambda x}$	λ = 0.05	Exponential	Low
C_{pareto}	$\alpha k^{\alpha} x^{-\alpha-1}$	α = 1.245, k = 3	Pareto	Medium
$C_{invgaus}$	$\sqrt{\frac{\lambda}{2\pi x^3}} e^{\frac{-\lambda(x-\mu)^2}{2\mu^2 x}}$	μ = 3.86, λ = 9.46	Inverse Gaussian	Medium
$C_{uniform}$	1	None	Uniform	High
$C_{lognorm}$	$\frac{1}{x\sqrt{2\pi\sigma^2}} e^{\frac{-(lnx-\mu)^2}{2\sigma^2}}$	μ = 5.929, σ = 0.321	Lognormal	High

PMF, probability mass function. $\delta(x)$, a PMF that returns 1 when $x = a$; otherwise returns 0.

In $\mathbb{A}_{E^2S}$, it is unnecessary to compute P_{opt}. P is set by $\mathbb{A}_{Init}$ and is indirectly referenced by $\mathbb{A}_{E^2S}$ through the set of partition blocks. P is an implicit and non-permutable value in $\mathbb{A}_{E^2S}$. Also, $\delta^{E^2S}_{part}$ is mainly bound to γ rather than to P.

5.4.2 Optimal Cost Tree for a Particular Number of Partition Blocks

For a fixed number of partition blocks P, the optimal number of K-ary cost tree K_{opt} is identified by searching K that produces minimum execution time from the following equation. $\delta^K_P(\mathbb{A}_{Init})$ is equation (7) that represents the execution time of the initial partitioning algorithm for a fixed P and an arbitrary K:

$$\min_{1<K\leq\infty}\{\delta^K_P(\mathbb{A}_{Init})\} \tag{20}$$

$$= \begin{cases} \min_{1<K<P}\{(3K+2)\left(\left\lceil\frac{P-K}{K-1}\right\rceil+1\right)+1\}, & 1<K<P \\ 3K+4, & K\geq P \end{cases}.$$

By rearranging equation (20), we obtain

$$\min_{1<K\leq\infty}\{\delta^K_P(\mathbb{A}_{Init})\} \tag{21}$$

$$= \begin{cases} \min_{1<K<P}\{5(P-1)\frac{1}{K-1}+3(P-1)+1\}, & 1<K<P \\ 3K+4, & K\geq P \end{cases}.$$

K_{opt} is $P-1$ when $1<K<P$ because $\delta^K_P(\mathbb{A}_{Init})$ decreases linearly as K increases. $\delta^{P-1}_P(\mathbb{A}_{Init})$ produces the minimum execution time. K_{opt} is P when $K\geq P$ because $\delta^K_P(\mathbb{A}_{Init})$ increases linearly as K increases:

$$K_{opt} = \begin{cases} P-1, & 1<K<P \\ P, & K\geq P \end{cases}. \tag{22}$$

In $\mathbb{A}_{E^2S}$, it is not necessary to compute K_{opt}. Once a model is selected by $\mathbb{A}_{Init}$, $\mathbb{A}_{E^2S}$ cannot dynamically permute structural properties of the model.

6. Experimental Results

A series of experiments have been conducted to evaluate the GMP algorithm and to compare it to two multiscale partitioning algorithms, *random* and *ratio-cut*. A set of quality and performance measures has been applied to the results. All experiments were performed on a small-scale Beowulf cluster system which consists of one root node, seven compute nodes, a Gigabit switch, and a dedicated Gigabit network. Each node is equipped with a single Intel Pentium 4 3 Ghz CPU, 2 GB PC 3200 DDR 400 SDRAM memory, a 80 GB 7200 RPM hard disk, and a Gigabit Ethernet card. The root node has an extra Gigabit Ethernet card to access public Internet.

A multiscale decomposable DEVS coupled model is generated with a particular cost pattern listed in Table 2. Six cost patterns are used to represent a wide range of computational workload of the model: $C_{unitstep}$ and C_{exp} for low workload, C_{pareto} and C_{invgau} for medium workload, and $C_{uniform}$ and $C_{lognormal}$ for high workload. Specifically, the model is characterized by $T(d,k,n)$ as discussed in Section 5.2. Simulation activity of each atomic component is assigned based on one of the workload patterns. Each atomic component executes its proactive temporal behavior that is bound to the given activity value. Simulation activity of a coupled component is the aggregation of activities of its child components. There exists no I/O exchange between components. The generated coupled model is partitioned into a set of partition blocks by each algorithm. The blocks are then dispatched to a set of processors, and decomposed models in the blocks are executed in parallel. Quality and performance of algorithms are computed by applying qualitative and quantitative measures to the blocks. We conducted experiments using the simulation environment that we developed for this research.

A set of measures used in our experiments is listed in Table 3. The cost of a partition block is computed by four cost measures: ϕ_{norm}, ϕ_{diff}, ϕ_{dist}, and ϕ_{var}. The quality of a set of partition blocks is evaluated by two QoP measures: $\varphi_{\min-\max}$ and $\varphi_{avg\text{-}diff}$. QoP evolution is traced by profiling the quality of the set until the best partitioning

Table 3. A set of measures for quality and performance evaluation

Measure	Mathematical representation	Description
ϕ_{norm}	$\mathrm{f}_{aggr}(PB_i)/\max\{\mathrm{f}_{aggr}(PB_j)\}_{j=1}^{\mathrm{P}}$	normalized cost
ϕ_{diff}	$\sum_{j=1}^{\mathrm{P}} \lvert f_{aggr}(PB_i) - f_{aggr}(PB_j)\rvert$	cost difference
ϕ_{dist}	$\sum_{j=1}^{\mathrm{P}} \lvert f_{aggr}(PB_i) - f_{aggr}(PB_j)\rvert$	cost distance
ϕ_{var}	$\sqrt{\lvert f_{aggr}(\mathrm{PB}_i) - \sum_{j=1}^{\mathrm{P}} f_{aggr}(PB_j)/\mathrm{P}\rvert}$	cost variance
$\varphi_{\min\text{-}\max}$	$\lvert \max\{f_{aggr}(PB_i)\}_{i=1}^{\mathrm{P}} - \min\{f_{aggr}(PB_i)\}_{i=1}^{\mathrm{P}}\rvert$	min-max disparity
$\varphi_{avg\text{-}diff}$	$\sum_{i=1}^{\mathrm{P}}\sum_{j=1}^{\mathrm{P}} \lvert f_{aggr}(PB_i) - f_{aggr}(PB_j)\rvert/P$	average difference
δ_{total}	$\max\{\tau(\mathrm{PB}_i)\}_{i=1}^{\mathrm{P}}$	total execution time
δ_{accum}	$\sum_{i=1}^{\mathrm{P}} \tau(PB_i)$	accumulated execution time
δ_{avg}	$\sum_{i=1}^{\mathrm{P}} \tau(PB_i)/P$	average execution time
δ_{sqrt}	$\sqrt{\sum_{i=1}^{\mathrm{P}} \tau(\mathrm{PB}_i)/\mathrm{P}}$	square root execution time

$f_{aggr}(\mathrm{PB}_i) = \sum c(t_j)$, where $c(t_j)$ is the cost of a model $t_j \in \mathrm{PB}_i$; $\tau(\mathrm{PB}_i)$ = time need to execute all models in PB_i.

result is attained. Execution time of the set is collected by four time measures: δ_{total}, δ_{accum}, δ_{avg}, and δ_{sqrt}.

Figure 4 represents the QoP experimental results of $T(7, 4, 400)$. Two QoP measures $\varphi_{\min-\max}$ and $\varphi_{avg-diff}$ are used to evaluate the quality of the partitioning results produced by each algorithm. For a particular number of partition blocks P, a cost pattern $C_{pattern}$, and a QoP measure, each algorithm is applied to 20 different computational workloads. The average of the 20 experiments is computed and illustrated as a single point in the figure. We measured QoP of partitioning results by varying P from 2 to 100. The GMP algorithm produces superior QoP outcomes compared to other algorithms for both QoP measures. This is mainly because the GMP algorithm minimizes the cost disparities between partition blocks with minimum model decomposition. However, QoP outcomes of other algorithms were highly sensitive to P, $C_{pattern}$, and QoP measures. The QoP experimental results are shown in Figure 4. All experimental results in the figure are summarized in Table 4.

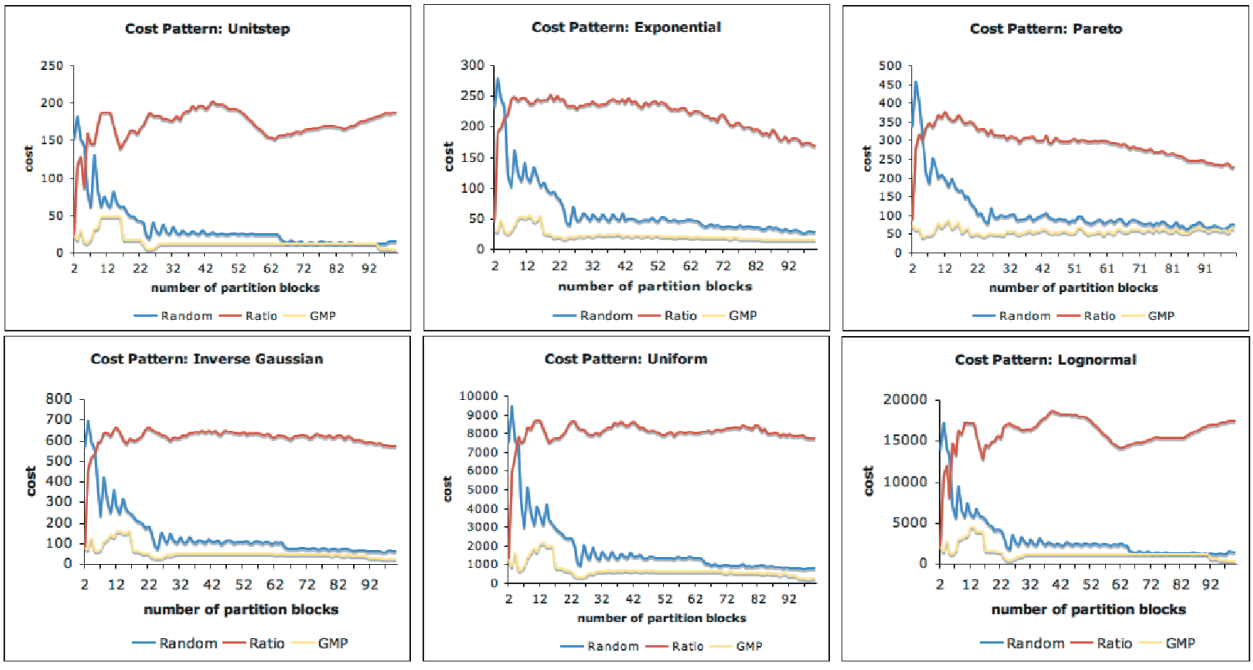
Figure 5 represents the execution time measurement results of $T(7, 4, 50)$. Two time measures, δ_{accum} and δ_{avg}, are used to evaluate the performance of the partitioning results produced by each algorithm. The model execution time is measured for the case of only one partition block allocation per processor. For a particular number of processors np and $C_{pattern}$, and a time measure, each algorithm is applied to five different computational workloads. In most experiments, the GMP algorithm requires the shortest execution time. The execution time measurement results are shown in Figure 5. All experimental results in the figure are summarized in Table 5.

7. Summary and Conclusions

In this paper we have presented a new GMP algorithm, which efficiently decomposes a multiscale model into a set of partition blocks using the cost modeling and analysis method. It also produces monotonically improved partitioning results with minimum model decomposition. The method enables abstracting subjective, heterogeneous, domain-dependent information into objective, homogeneous, domain-independent cost information. With the selection of different methods of cost measures, cost evaluation, and cost aggregation, the proposed algorithm performs various partitioning strategies without any modification of the generic partitioning logics. Because each cost measure is a parametric method, and partitioning logic is described over the homogeneous cost space, the algorithm is generic and applicable to any family of models provided that there is a way to manipulate the appropriate cost information.

Algorithm analysis and experimental results show that the GMP algorithm is efficient and produces high-quality partitioning results. The algorithm execution time is $O(n)$

a) QoP measure: min-max disparity



b) QoP measure: average difference

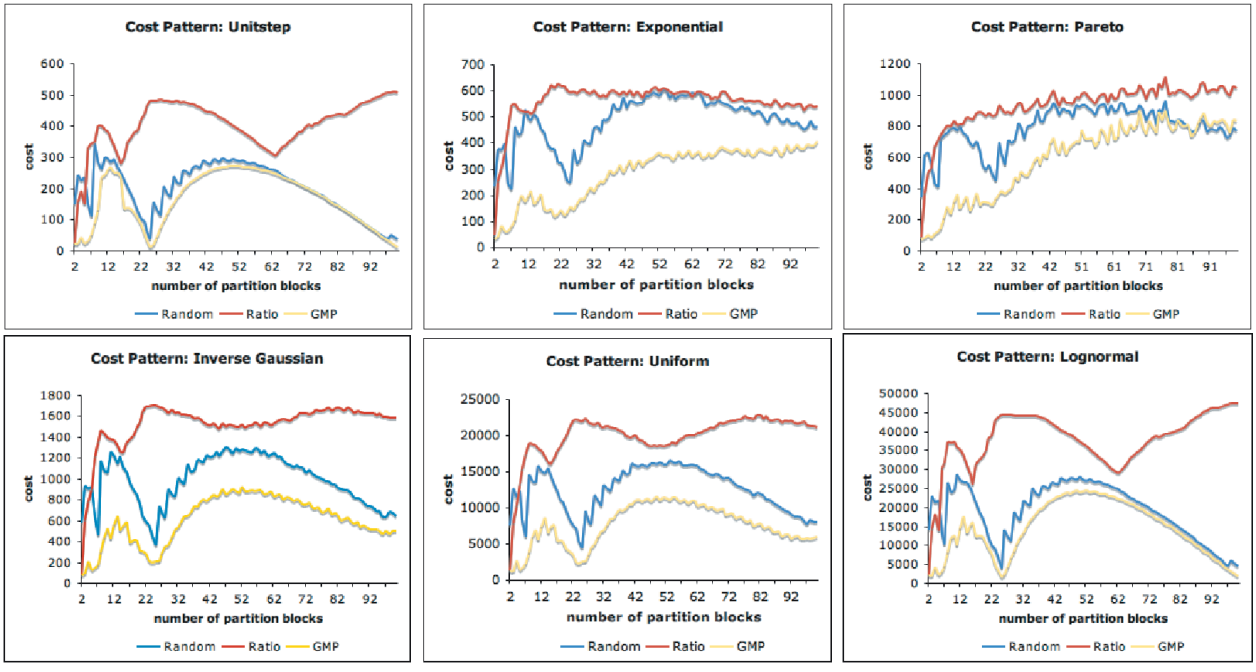
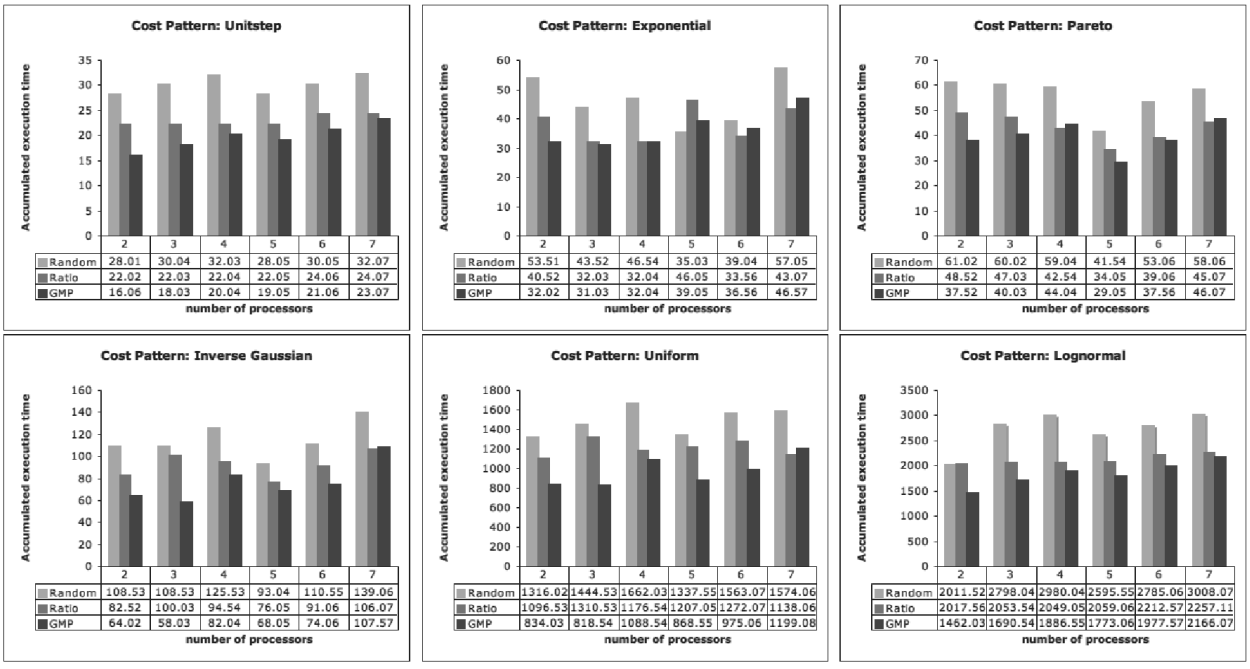


Figure 4. QoP evaluation of partitioning results over various cost patterns and numbers of partition blocks. A point in each figure represents the average of 20 different executions with respect to a set of a cost pattern, a number of partition blocks, and a partitioning algorithm, $C_{pattern}$, P, $\mathbb{A}_{partition}$. $C_{pattern} \in \{C_{unitstep}, C_{exp}, C_{pareto}, C_{invgaus}, C_{uniform}, C_{lognorm}\}$. $2 \leq P \leq 100$. $\mathbb{A}_{partition} \in \{Random, Ratio\text{-}Cut, GMP\}$. Two QoP measures, $\varphi_{min-max}$ and $\varphi_{avg-diff}$, are applied to $T(7, 4, 400)$. The lower value on the Y-axis represents the better result.

a) Execution time measure: accumulated execution time



b) Execution time measure: average execution time

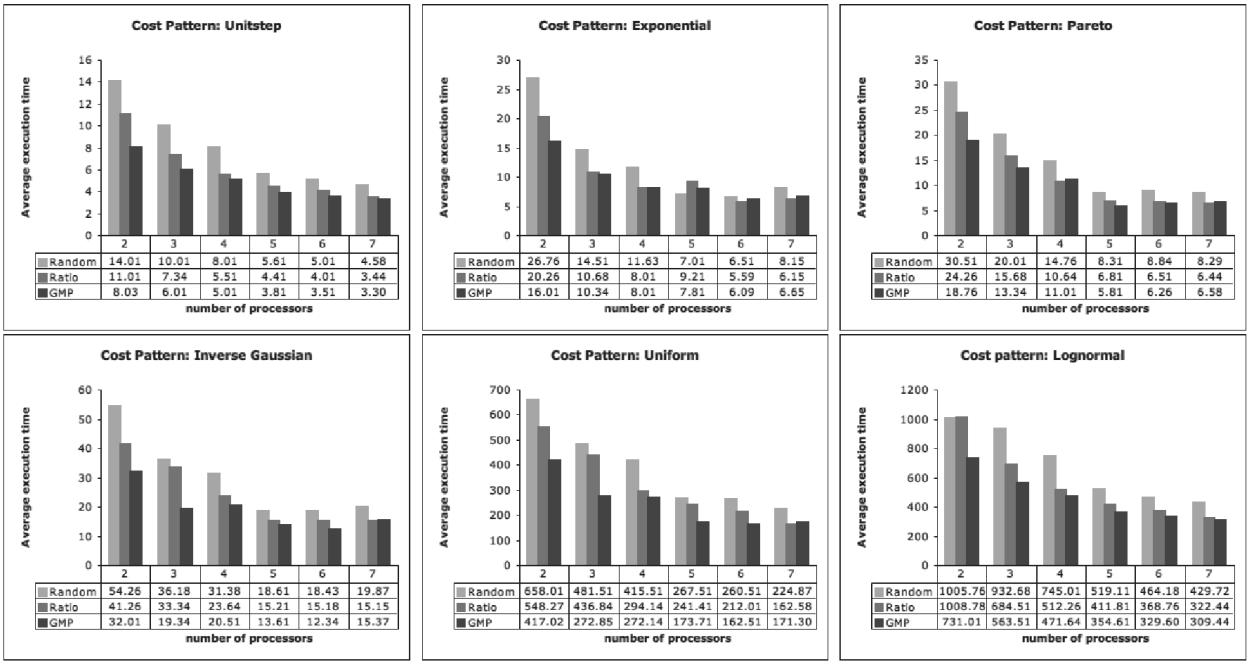


Figure 5. Model execution time measurement of partitioning results over various cost patterns and numbers of processors. Each mark in the figure is the average of five different executions with respect to a set of a cost pattern, a number of processors, and a partitioning algorithm, $C_{pattern}, np, \mathbb{A}_{partition}$. $C_{pattern} \in \{C_{unitstep}, C_{exp}, C_{pareto}, C_{invgaus}, C_{uniform}, C_{lognorm}\}$. $2 \leq np \leq 7$. $\mathbb{A}_{partition} \in \{Random, Ratio-Cut, GMP\}$. Two execution time measures, δ_{accum} and δ_{avg}, are applied to $T(7, 4, 50)$. The lower value on the Y-axis represents the better result.

Table 4. A summary of QoP experiments of $T(7, 4, 400)$

	$\varphi_{\min-\max}$			$\varphi_{avg-diff}$		
Cost pattern	Random (s)	Ratio-cut (s)	GMP (s)	Random (s)	Ratio-cut (s)	GMP (s)
$C_{unitstep}(x)$	33.44	170.28	14.77	198.79	402.94	165.11
$C_{exp}(x)$	62.18	218.27	22.57	484.83	559.83	284.50
$C_{pareto}(x)$	110.44	292.09	56.84	788.20	929.16	595.32
$C_{invgaus}(x)$	139.83	609.87	53.60	984.37	1520.27	603.98
$C_{uniform}(x)$	1799.33	8000.09	689.31	12492.31	19945.52	7521.78
$C_{lognorm}(x)$	3077.86	15873.68	1303.46	18884.30	37774.36	14240.95
Average	870.51	4194.05	356.76	5638.80	10188.68	3901.94

Table 5. The summary of execution time measurements of $T(7, 4, 50)$

	δ_{accum}			δ_{avg}		
Cost pattern	Random (s)	Ratio-cut (s)	GMP (s)	Random (s)	Ratio-cut (s)	GMP (s)
$C_{unitstep}(x)$	30.04	22.71	19.55	7.87	5.95	4.94
$C_{exp}(x)$	45.78	37.88	36.21	12.43	9.98	9.15
$C_{pareto}(x)$	55.45	42.71	39.04	15.12	11.72	10.29
$C_{invgaus}(x)$	114.21	91.71	75.63	29.79	23.96	18.86
$C_{uniform}(x)$	1482.88	1200.13	963.97	384.65	315.87	244.92
$C_{lognorm}(x)$	2696.38	2108.15	1825.97	682.74	551.43	459.97
Average	737.46	583.88	493.39	188.77	153.15	124.69

in the worst-case scenario. The experimental results show that the GMP algorithm produces partitioning results that are superior to those from other algorithms.

A set of GMP-based multiscale model partitioners has been implemented over distributed network middleware to support large-scale discrete-event oriented simulations. Because the algorithm is generic, concise, and reconfigurable, it can easily evolve to accommodate static and dynamic resource management system components that efficiently handle multiscale models in large-scale distributed and parallel simulation systems.

The pace of M&S driven systems biology research using constructive, multiscale models is expected to increase. However, for efficient execution of these models, we need generic but domain-aware, multiscale partitioning algorithms. The GMP algorithm meets this requirement and has been successfully implemented as a part of multiscale model partitioners in various large-scale distributed simulation frameworks. A wide range of distinctive multiscale, constructive, modular biological system models can be easily managed by changing or revising cost functions without any modification of generic partitioning programming logics. This ability positions the GMP algorithm to be effective in large-scale M&S driven systems biology research. We anticipate applying the GMP algorithm in our computational systems biology research.

Acknowledgments

The authors would like to thank Sean H. J. Kim, Dr James Nutaro, Dr Hessam Sarjoughian, and the anonymous reviewers for their constructive comments and suggestions that helped improve the content of the paper. This research has been supported in part by the CDH Research Foundation R21-CDH-00101, NSF DMI-0122227, and DOE SciDAC DE-FC02-01ER41184. We are grateful for the Computational and Systems Biology Postdoctoral Fellowship funding provided to Sunwoo Park by the CDH Foundation. A preliminary version of this paper was presented at the Challenges of Large Applications in Distributed Environments (CLADE) 2003, International Workshop on Heterogeneous and Adaptive Computation, June 2003.

Appendix

A. Length of the component list $clist$ after i node expansions l_i

As described in Algorithm 1, $clist$ is initially populated with child nodes of the root node in a cost tree. Thus, the initial length of the $clist$ l_0 is equivalent to the number of the nodes ξ_0. When a node expansion occurs, a coupled node is removed from the $clist$ and its child nodes are stored back to the $clist$. That is, the length of the $clist$ at the ith node expansion l_i is equivalent to $l_{i-1} + \xi_i$ where $i \geq 1$ and ξ_i is the number of child nodes of the removed coupled node at the ith node expansion. Given ξ_0 and ξ_i, l_i is deduced as follows:

$$
\begin{aligned}
l_0 &= \xi_0 \\
l_1 &= l_0 - 1 + \xi_1 = \xi_0 + \xi_1 - 1 \\
l_2 &= l_1 - 1 + \xi_2 = (\xi_0 + \xi_1 - 1) - 1 + \xi_2 \\
&= \xi_0 + \xi_1 + \xi_2 - 2 \\
l_3 &= l_2 - 1 + \xi_3 = (\xi_0 + \xi_1 + \xi_2 - 2) - 1 + \xi_3 \\
&= \xi_0 + \xi_1 + \xi_2 + \xi_3 - 3 \\
&\cdots \\
l_{i-1} &= l_{i-2} - 1 + \xi_{i-1} \\
&= (\xi_0 + \cdots + \xi_{i-2} - (i-2)) - 1 + \xi_{i-1} \\
&= \xi_0 + \cdots + \xi_{i-1} - (i-1) \\
l_i &= l_{i-1} - 1 + \xi_i \\
&= (\xi_0 + \cdots + \xi_{i-1} - (i-1)) - 1 + \xi_i \\
&= \xi_0 + \cdots + \xi_i - i.
\end{aligned}
$$

By replacing l_{i-1} by the sum of ξ up to $(i-1)$th node expansions, we rewrite l_i as follows:

$$
\begin{aligned}
l_i &= l_{i-1} - 1 + \xi_i = \sum_{j=0}^{i-1} (\xi_j - 1) - 1 + \xi_i \\
&= \left(\sum_{j=0}^{i-1} \xi_j + \xi_i \right) - \left(\sum_{j=0}^{i-1} 1 + 1 \right) \\
&= \sum_{j=0}^{i} \xi_j - i, \quad i \geq 1. \qquad (23)
\end{aligned}
$$

Assume the initial partitioning requires E expansions to guarantee that a partition has at least one cost node. Then, by applying equation (23) to the *while* loop (line 6, Algorithm 1) with substitution of E for i, we assert the constraint using E, ξ_j, and P as follows in the partitioning algorithm, where P is the number of partition blocks:

$$
\begin{aligned}
lengthOf(clist) &< numberOf(parray) \\
&= \left(\sum_{j=0}^{E} \xi_j - E \right) < P. \qquad (24)
\end{aligned}
$$

B. Total number of node expansions in the initial partitioning E

To make analysis simple, let ξ_i be a constant K (i.e. K-ary tree), which implies that every coupled node has only K child nodes. No expansion occurs when $K \leq P$ and E expansions occur when $1 < K < P$. By substituting K for ξ_i in equation (24), we obtain

$$
\begin{aligned}
\left(\sum_{j=0}^{E} K - E \right) &< P = ((E+1)K - E) < P \\
&= (K-1)E < (P-K) \\
&= E < \frac{P-K}{K-1}, \quad 1 < K < P, \qquad (25)
\end{aligned}
$$

provided that P and K, the total number of node expansions needed in the initial partitioning, is represented by

$$
E = \begin{cases} \left\lceil \frac{P-K}{K-1} \right\rceil. & 1 < K < P \\ 0. & K \geq P \end{cases}. \qquad (26)
$$

8. References

[1] Kitano, H. 2002. Computational systems biology. *Nature* 420:206–210.

[2] Idekeer, T., T. Galitski, and L. Hood. 2001. A new approach to decoding life: Systems biology. *Annual Review of Genomics and Human Genetics* 2:343–372.

[3] Hood, L. 2003. Systems biology: Integrating technology, biology, and computation. *Mechanisms of Ageing and Development* 124(1):9–16.

[4] Park, S. 2003. Cost-based partitioning for distributed simulation of hierarchical, modular DEVS models. PhD Dissertation. Department of Electrical and Computer Engineering, University of Arizona.

[5] Park, S., C. A. Hunt, and G. E. P. Ropella. 2005. Pisl: A large-scale in silico experimentation framework for agent-directed physiological models. In *Proceedings of the 2005 Agent-Directed Simulation Symposium, Spring Simulation Conference*, San Diego, CA.

[6] Zeigler, B. P., H. Praehofer, and T. G. Kim. 2000. *Theory of Modeling and Simulation*, 2nd edn. San Diego, CA: Academic Press.

[7] Vangheluwe, H. L. 2000. DEVS as a common denominator for multi-formalism hybrid systems modeling. In *Proceedings of the 2000 IEEE International Symposium on Computer-Aided Control System Design*, Anchorage, AK.

[8] Zeigler, B. P. 2003. DEVS today: Recent advances in discrete event-based information technology. In *Proceedings of the 11th IEEE/ACM International Symposium on Modeling, Analysis and Simulation of Computer Telecommunications Systems (MASCOTS 2003)*, Orlando, FL, pp. 141–161.

[9] Djafarzadeh, R., G. Waineer, and T. Mussivand. 2005. DEVS modeling and simulation of the cellular metabolism by mitochondria. In *Proceedings of the 2005 DEVS Integrative M&S Symposium, Spring Simulation Conference*, San Diego, CA, pp. 55–62.

[10] Hu, X., and D. H. Edwards. 2005. Behaviorsim: A simulation environment to study animal behavioral choice mechanisms. in *Proceedings of the 2005 DEVS Integrative M&S Symposium, Spring Simulation Conference*, San Diego, CA.

[11] Biermann, S., A. M. Uhrmacher, and H. Schumann. 2004. Supporting multi-level models in systems biology by visual methods. In *Proceedings of the 18th European Simulation Multiconference*, Magdeburg, Germany.

[12] Uhrmacher, A. M., D. Degenring, and B. P. Zeigler. 2005. Discrete event multi-level models for systems biology. *Lecture Notes in Computer Science* Vol. 3380. Berlin: Springer, pp. 66–89.

[13] Seo, C., S. Park, B. Kim, S. Cheon, and B. P. Zeigler. 2004. Implementation of distributed high-performance devs simulation framework in the grid computing environment. In *Proceedings of the 2004 Advanced Simulation Technologies Conference (ASTC '04) – High Performance Computing Symposium 2004 (HPCS 2004)*, Arlington, VA.

[14] Cheon, S., C. Seo, S. Park, and B. P. Zeigler. 2004. Design and implementation of distributed DEVS simulation in a peer-to-peer network system. In *Proceedings of the 2004 Advanced Simulation Technologies Conference (ACTS '04) – Design, Analysis, and Simulation of Distributed Systems Symposium 2004 (DASD 2004)*, Arlington, VA.

[15] Pothen, A. 1996. Graph partitioning algorithms with applications to scientific computing. In *Parallel Numerical Algorithms*, D. E. Keyes, A. Sameh, and V. Venkatakrishnan, eds. Dordrecht: Kluwer, pp. 323–368.

[16] Fjällström, P.-O. 1998. Algorithms for graph partitioning: A survey. *Linkköping Electronic Articles in Computer and Information Science* 3(10).

[17] Alpert, C. J., and A. B. Kahng. 1995. Recent directions in netlist partitioning: A survey. *SIAM Journal on Scientific Computing* 16(1–2):452–469.

[18] Schloegel, K., G. Karypis, and V. Kumar. 2004. Graph partitioning for high performance scientific simulations. *Computing Reviews* 45(2).

[19] Karypis, G., and V. Kumar. 1998. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on Scientific Computing* 20:359–92.

[20] Barnard, S. T., and H. D. Simon. 1994. A fast multilevel implementation of recursive spectral bisection for partitioning unstructured problems. *Concurrency: Practice and Experience* 6:101–17.

[21] Saab, Y. G. 2004. An effective multilevel algorithms for bisecting graphs and hypergraphs. *IEEE Transactions on Computers* 53(6):641–52.

[22] Möller, M. O., and R. Alur. 2001. Heuristics for hierarchical partitioning with application to model checking. In *Proceedings of the 11th IFIP WG 10.5 Advanced Research Working Conference on Correct Hardware Design and Verification Methods*, Livingston, Scotland, pp. 71–85.

[23] Walshaw, C. 2004. Multilevel refinement for combinatorial optimisation problems. *Annals of Operations Research* 131(1–4):325–72.

[24] Kim, K., T. Kim, and K. Park. 1998. Hierarchical partitioning algorithm for optimistic distributed simulation of DEVS models. *Journal of Systems Architecture* 44:433–55.

[25] Zhang, G., and B. P. Zeigler. 1990. Mapping hierarchical discrete models to multiprocessor system: Concept, algorithm, and simulation. *Journal of Parallel and Distributed Computing* 9:271–80.

[26] Kernighan, B., and S. Lin. 1970. An efficient heuristic procedure for partitioning graphs. *The Bell Technical Journal* 49:291–307.

[27] Dutt, S. 1993. New faster Kernighan–Lin-type graph-partitioning algorithms. In *Proceedings of the 1993 IEEE/ACM International Conference on Computer-aided Design*, Santa Clara, CA, pp. 370–7.

[28] Fiduccia, C. M., and R. M. Mattheyses. 1982. A linear-time heuristic for improving network partitions. In *Proceedings of the IEEE/ACM 19th Design Automation Conference*, Las Vegas, NV, pp. 175–81.

[29] Diekmann, R., B. Monien, and R. Preis. 1994. Using helpful sets to improve graph bisections. Technical Report TR-RF-94-008, Department of Computer Science, University of Paderborn, Germany.

[30] Gilbert, J. R., and E. Zmijewski. 1987. A parallel graph partitioning algorithm for a message-passing multiprocessor. *International Journal of Parallel Programming* 16(6):427–49.

[31] Berger, M. J., and B. H. Bokhari. 1987. A partitioning strategy for non-uniform problems across multiprocessors. *IEEE Transactions on Computers* 36:570–80.

[32] Williams, R. D. 1991. Performance of dynamic load balancing algorithms for unstructured mesh calculations. *Concurrency: Practice and Experience* 3:457–81.

[33] Fox, G. C. 1988. *A Review of Automatic Load Balancing and Decomposition Methods for the Hypercube*. New York: Springer-Verlag, pp. 63–76.

[34] Farhat, C., and M. Lesoinne. 1993. Automatic partitioning of unstructured meshes for the parallel solutions of problems in computational mechanisms. *International Journal for Numerical Methods in Engineering* 36:745–64.

[35] Simon, H. D. 1991. Partitioning of unstructured problems for parallel processing. *Computing Systems in Engineering* 2(2–3):135–48.

[36] Pothen, A., H. D. Simon, and K. P. Liu. 1990. Partitioning sparse matrices with eigenvectors of graphs. *SIAM Journal on Matrix Analysis and Applications* 11(3):430–52.

[37] Barnes, E. R. 1982. An algorithm for partitioning the nodes of a graph. *SIAM Journal for Algorithms and Discrete Methods* 3:541–50.

[38] Boppana, R. B. 1987. An average case analysis. In *Proceedings of the 28th Annual IEEE Symposium on Foundations of Computer Science*, Los Angeles, CA, pp. 67–75.

[39] Hendrickson, B., and R. Leland. 1995. An improved spectral graph partitioning algorithm for mapping parallel computations. *SIAM Journal on Scientific Computing* 16(2):452–69.

[40] Banan, M., and K. D. Hielmstad. 1992. Self-organization of architecture by simulated hierarchical adaptive random partitioning. In *Proceedings of the International Joint Conference of Neural Networks (IJCNN)*, Vol. 3, pp. 823–8.

[41] Johnson, D. S., C. R. Aragon, L. A. McGeoch, and C. Schevon. 1989. Optimization by simulated annealing: an experimental evaluation. Part I: Graph partitioning. *Operations Research* 37(6):865–92.

[42] den Bout, D. E. V., and T. K. Miller, III. 1994. Graph partitioning using annealed neural networks. *IEEE transaction on Neural Networks* 1(2):192–203.

[43] Rolland, E., H. Pirkul, and F. Glover. 1996. Tabu search for graph partitioning. *Annals of Operations Research* 63:209–32.

[44] Bui, T. N., and B. R. Moon. 1996. Genetic algorithm and graph partitioning. *IEEE Transactions on Computers* 45(7):841–55.

[45] Cardellini, V., M. Colajanni, and P. S. Yu. 2003. Request redirection algorithms for distributed web systems. *IEEE Transactions on Parallel and Distributed Systems* 14(1):355–68.

[46] Floyd, S., and V. Paxson. 2001. Difficulties in simulating the internet. *IEEE/ACM Transactions on Networking* 9(4):392–403.

[47] Arlitt, M., and T. Jin. 2000. A workload characterization study of the 1998 World Cup web site. *IEEE Network* 14:30–7.
[48] Barford, P., and M. Crovella. 1999. A performance evaluation of hyper text transfer protocols. *Proceeding of ACM Sigmetrics* 27(1):188–97.

Sunwoo Park *is a postdoctoral fellow at the University of California at San Francisco, Department of Biopharmaceutical Sciences, San Francisco, California, USA.*

C. Anthony Hunt *is a Professor at the University of California at San Francisco, Department of Biopharmaceutical Sciences, San Francisco, California, USA.*

Bernard P. Zeigler *is a Professor at the University of Arizona at San Francisco, Department of Electrical and Computer Engineering, Tucson, Arizona, USA.*