

---

# Recent Research on the Cray XMT

**Shahid Bokhari**

Department of Biomedical Informatics

The Ohio State University

`shahid@bmi.osu.edu`

`bmi.osu.edu/~shahid`

`shb@acm.org`

`home.earthlink.net/~drshb`

November 17, 2009

# Example: Sequence Alignment

Ordinary C code—The compiler is reassured that P, T & D do not overlap: automatically converts rectangular loop into wavefronted loop.

```

|           #pragma mta noalias *P, *T, *D
|           for (i=1; i <=m; i++) {
|               int j;
7 P:e      |               int myPi=P[i];
|               for(j=1; j <= n ; j++){
|                   int v, h, d, m1, m2, p;
9 -P1:w    |                   v= D[i-1][j]+1;
|                   p= (myPi!=T[j]);
9 -P1:w    |                   ...
|                   m2 = MIN( m1,h);
9 P-:w     |                   D[i][j] = m2;
9 SP1:w    +
|               }
|           }

```

## Alignment

Comparison

Model

Reassortment

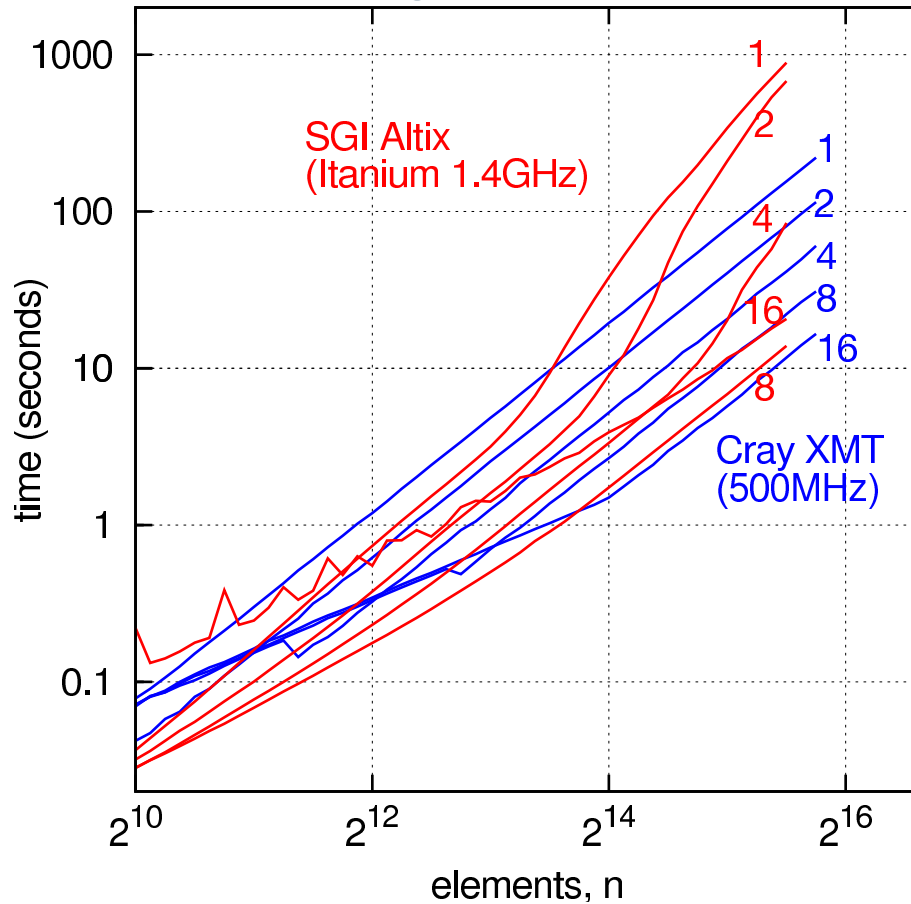
Network

Shared Multilevel

Parallel Queues

# Comparative Timings

## Sequence alignment on the Cray XMT & SGI Altix



### *Concurrency & Computation*

[www3.interscience.wiley.com/journal/  
122580907/abstract](http://www3.interscience.wiley.com/journal/122580907/abstract)

# Graph Theoretic Model for Virus Reassortment

Alignment

Comparison

Model

Reassortment

Network

Shared Multilevel

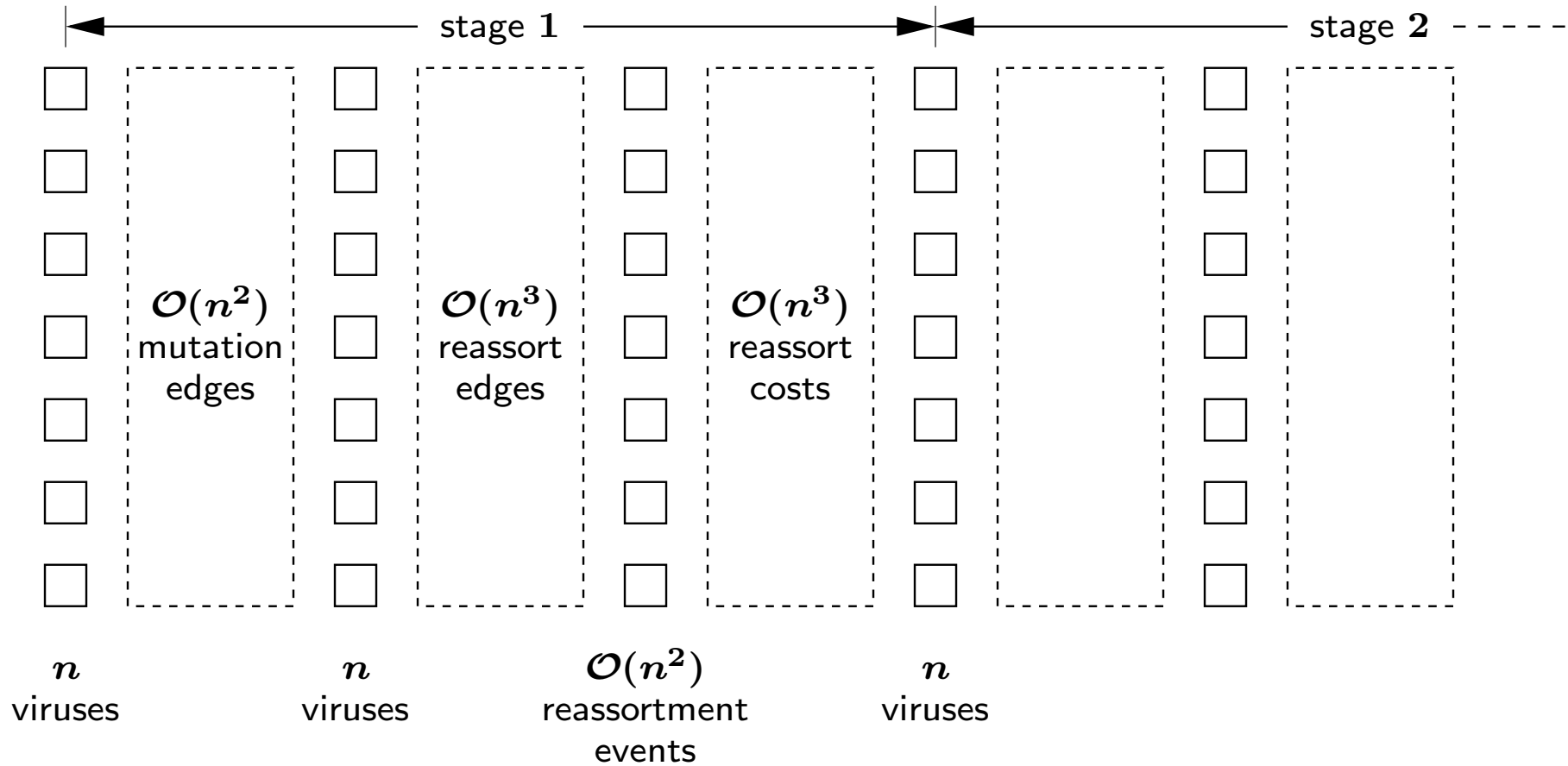
Parallel Queues

- Evolution over  $\tau$  stages or seasons modeled with a layered or *multipartite* graph.
- Viruses = nodes
- Reassortment events = nodes
- Mutations/reassortment choices = edges
- Weights on edges represent edit distances

*IEEE Trans. Computational Biology & Bioinformatics*

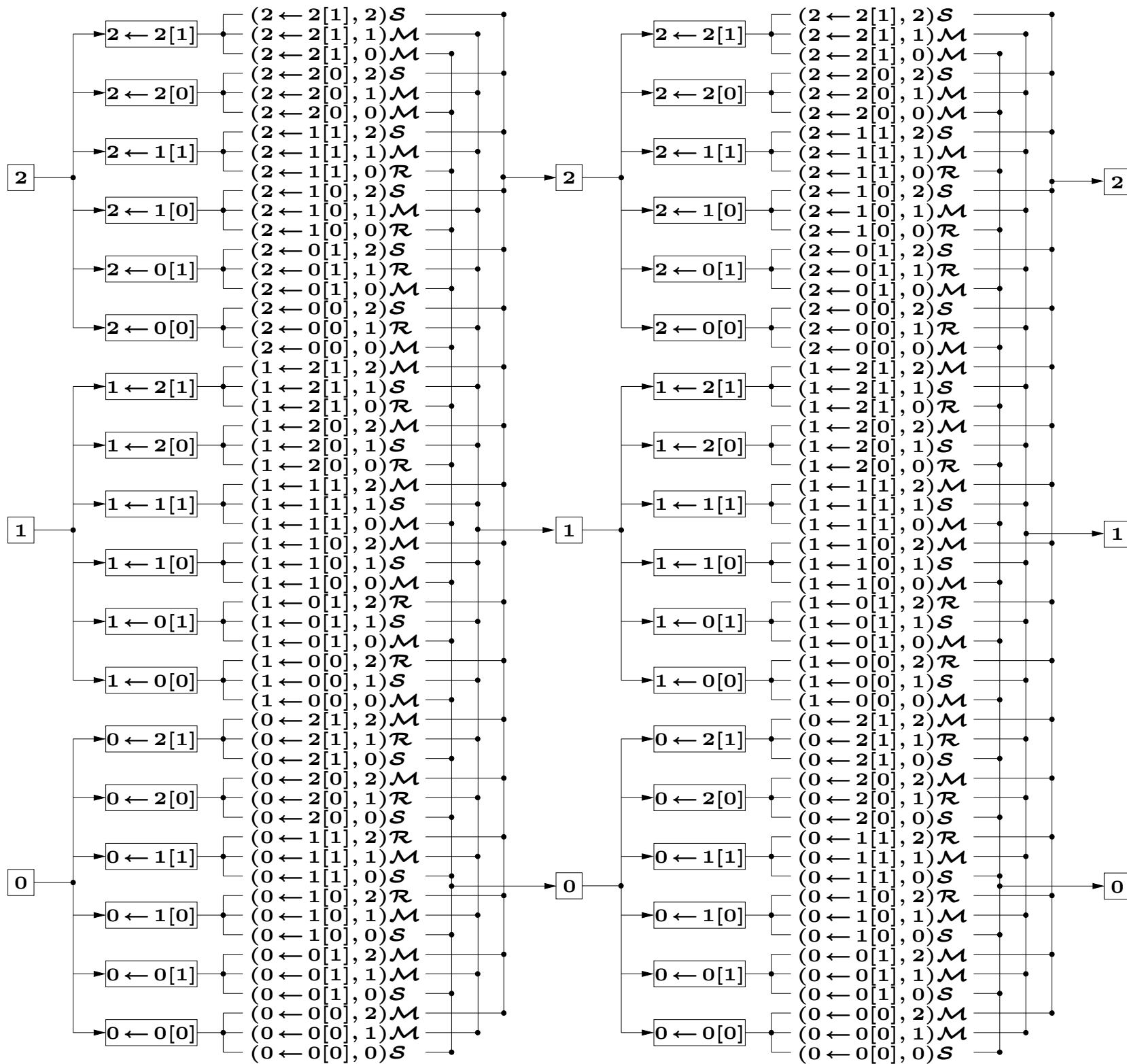
[doi.ieeecomputersociety.org/10.1109/TCBB.2008.73](https://doi.ieeecomputersociety.org/10.1109/TCBB.2008.73)

# Reassortment Network

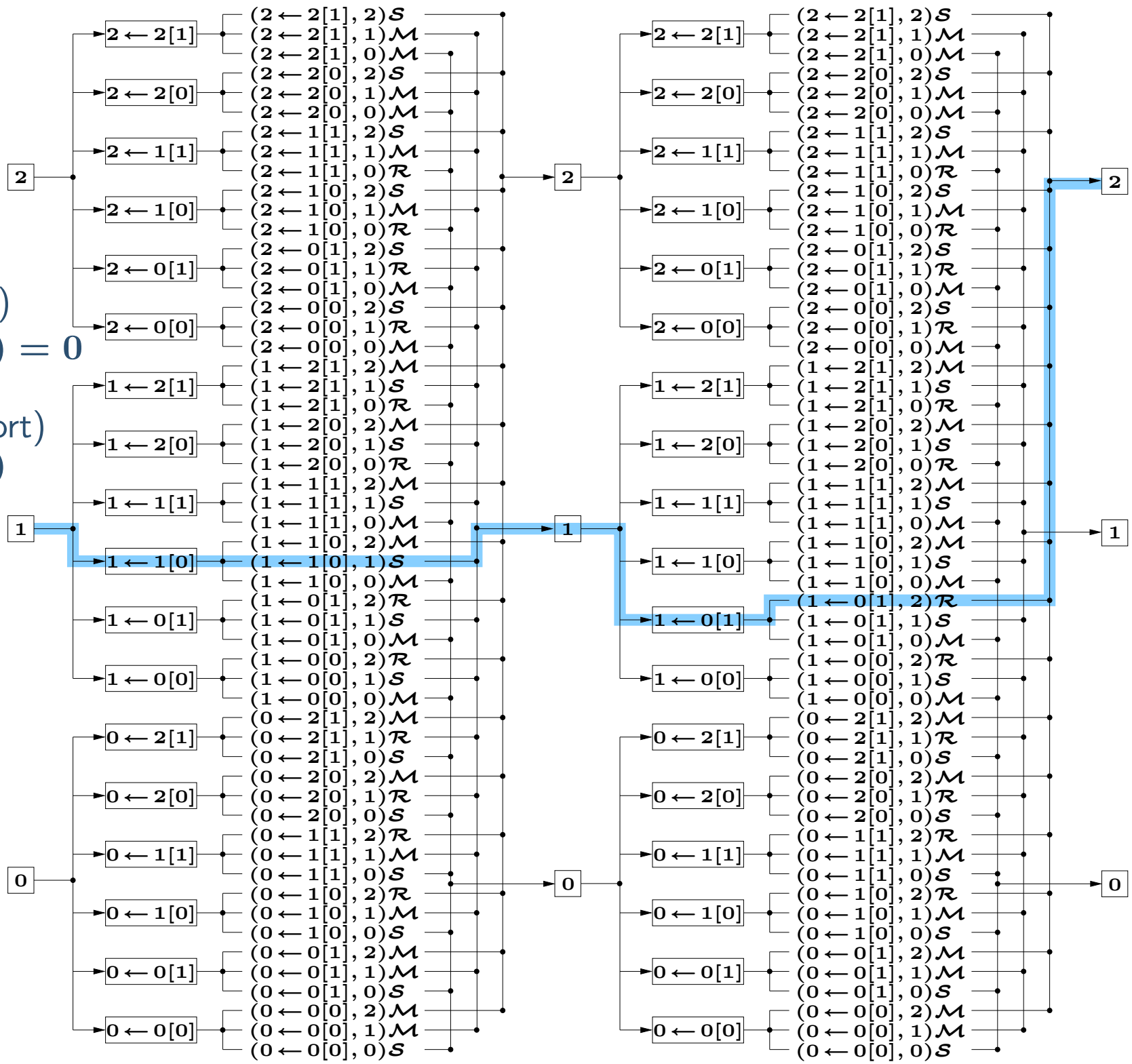


Path in Reassortment Network corresponds to sequences of mutations and reassortments that transform one virus into another.

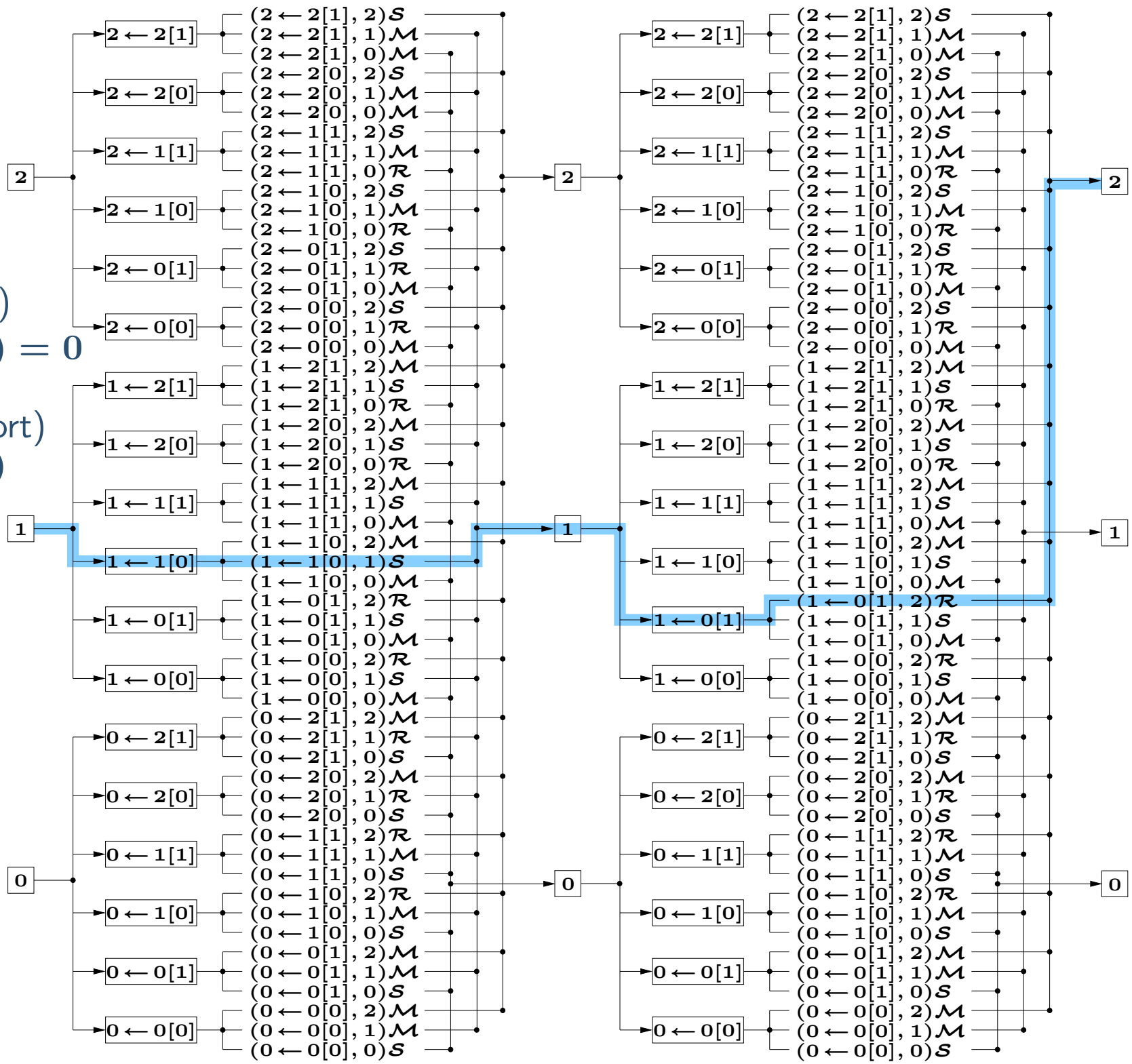
Network for  
 $n = 3$  viruses,  
 $\tau = 2$  stages,  
 2 segments/virus



Path:  
 Virus 1  
 $1 \leftarrow 1[0]$  (stasis)  
 $\mathcal{W}(1 \leftarrow 1[0], 1) = 0$   
 Virus 1  
 $1 \leftarrow 0[1]$  (reassort)  
 $\mathcal{W}(1 \leftarrow 0[1], 2)$   
 Virus 2

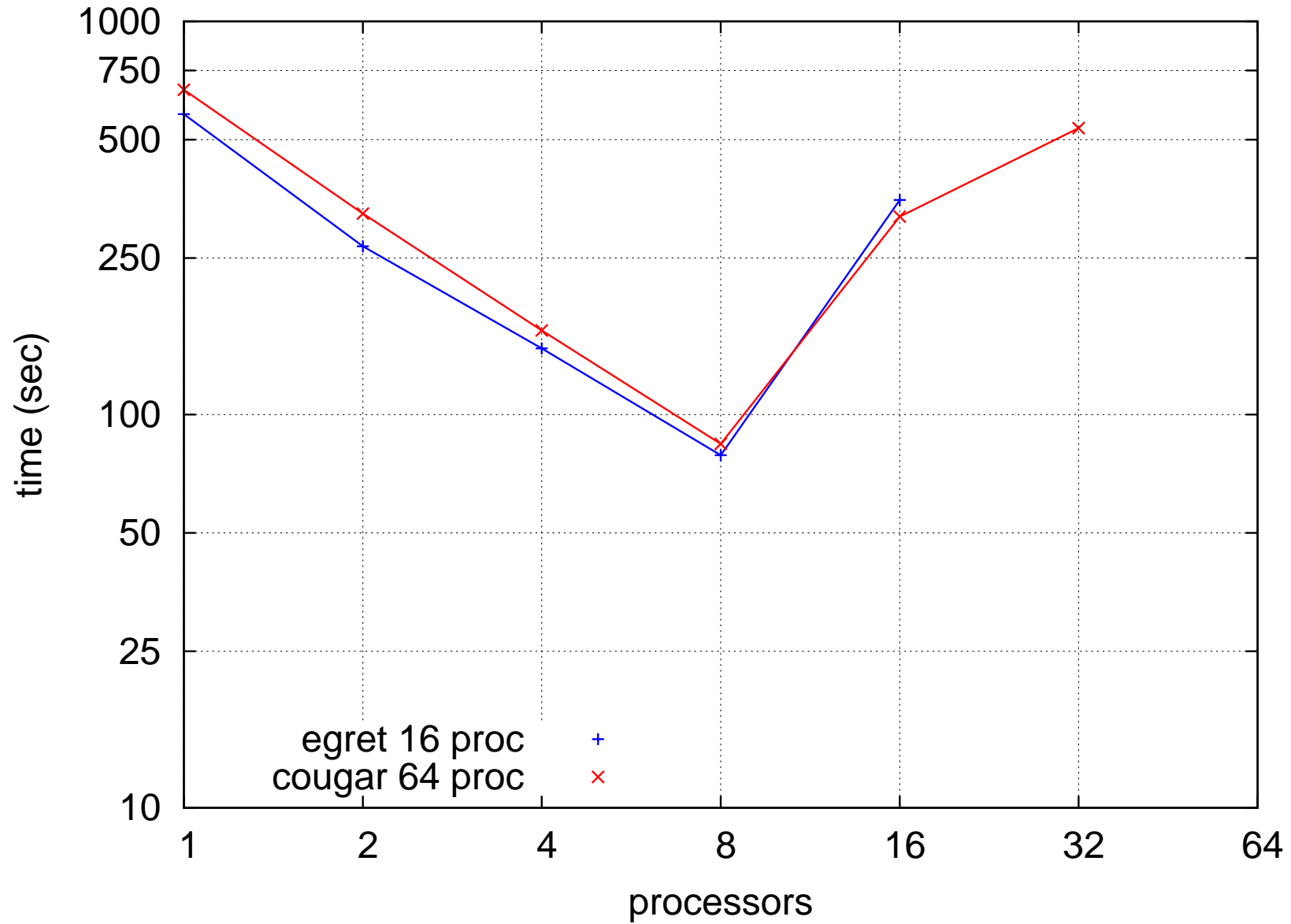


Path:  
 Virus 1  
 $1 \leftarrow 1[0]$  (stasis)  
 $\mathcal{W}(1 \leftarrow 1[0], 1) = 0$   
 Virus 1  
 $1 \leftarrow 0[1]$  (reassort)  
 $\mathcal{W}(1 \leftarrow 0[1], 2)$   
 Virus 2

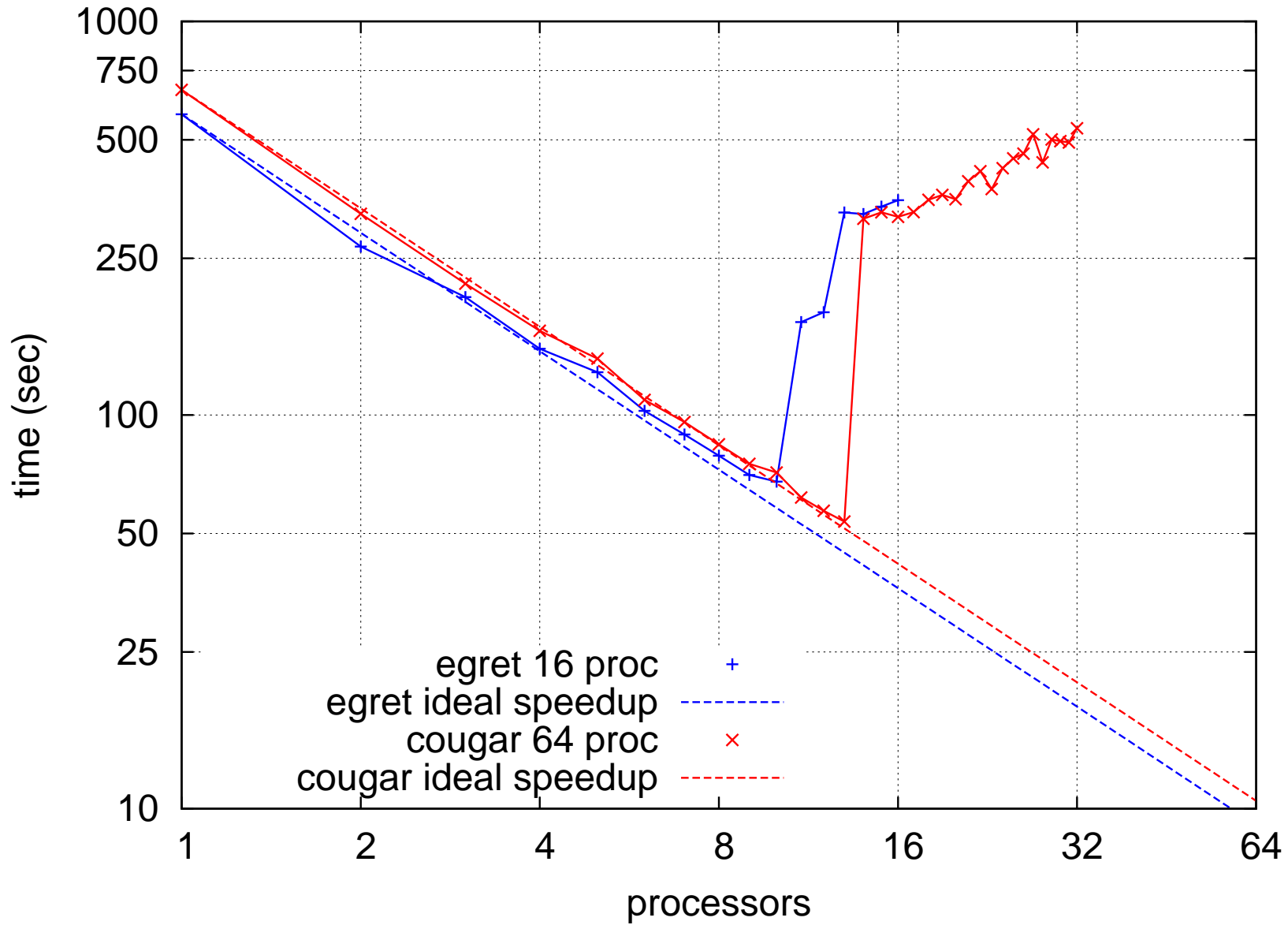




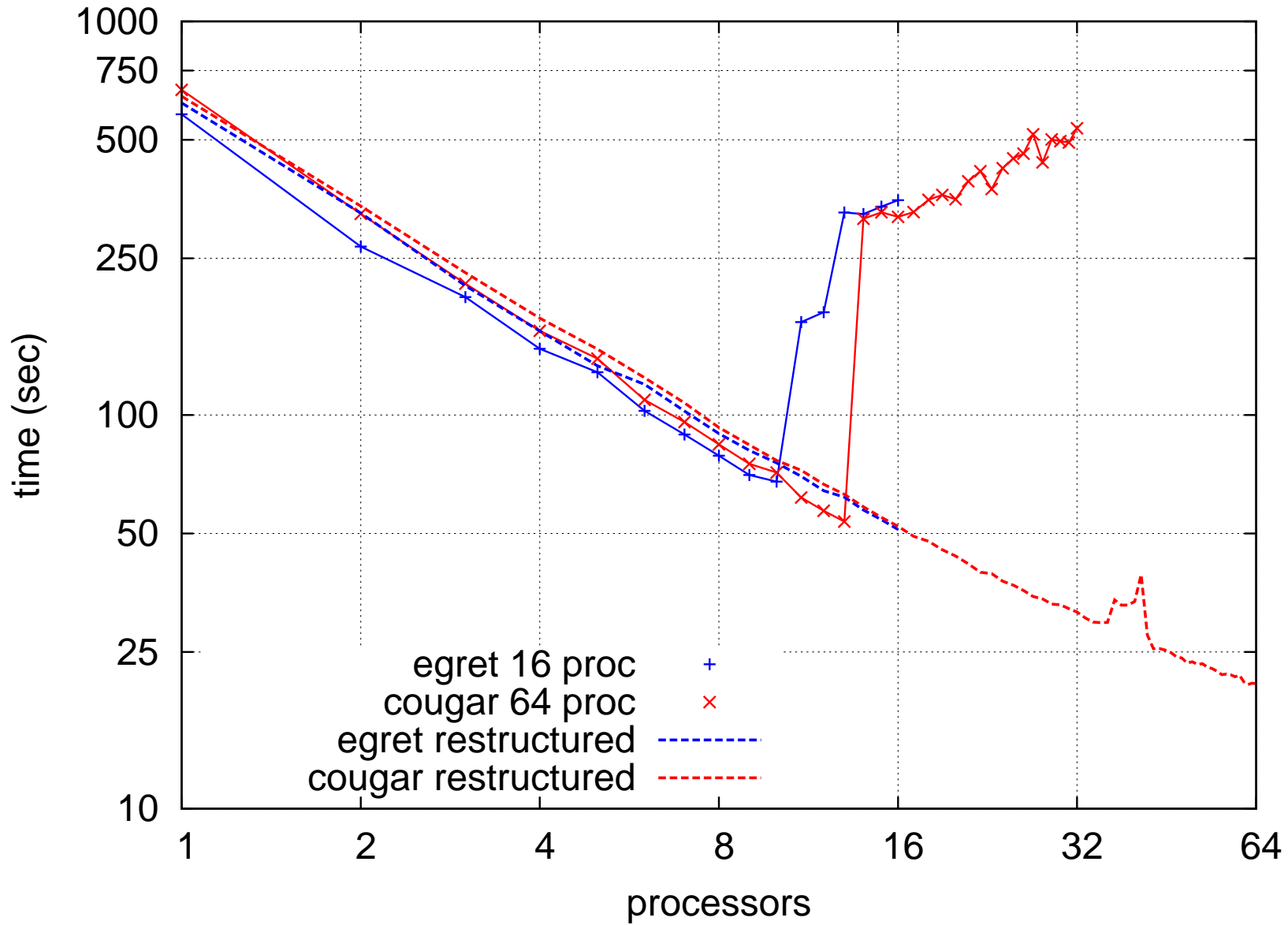
# Comparative Timings



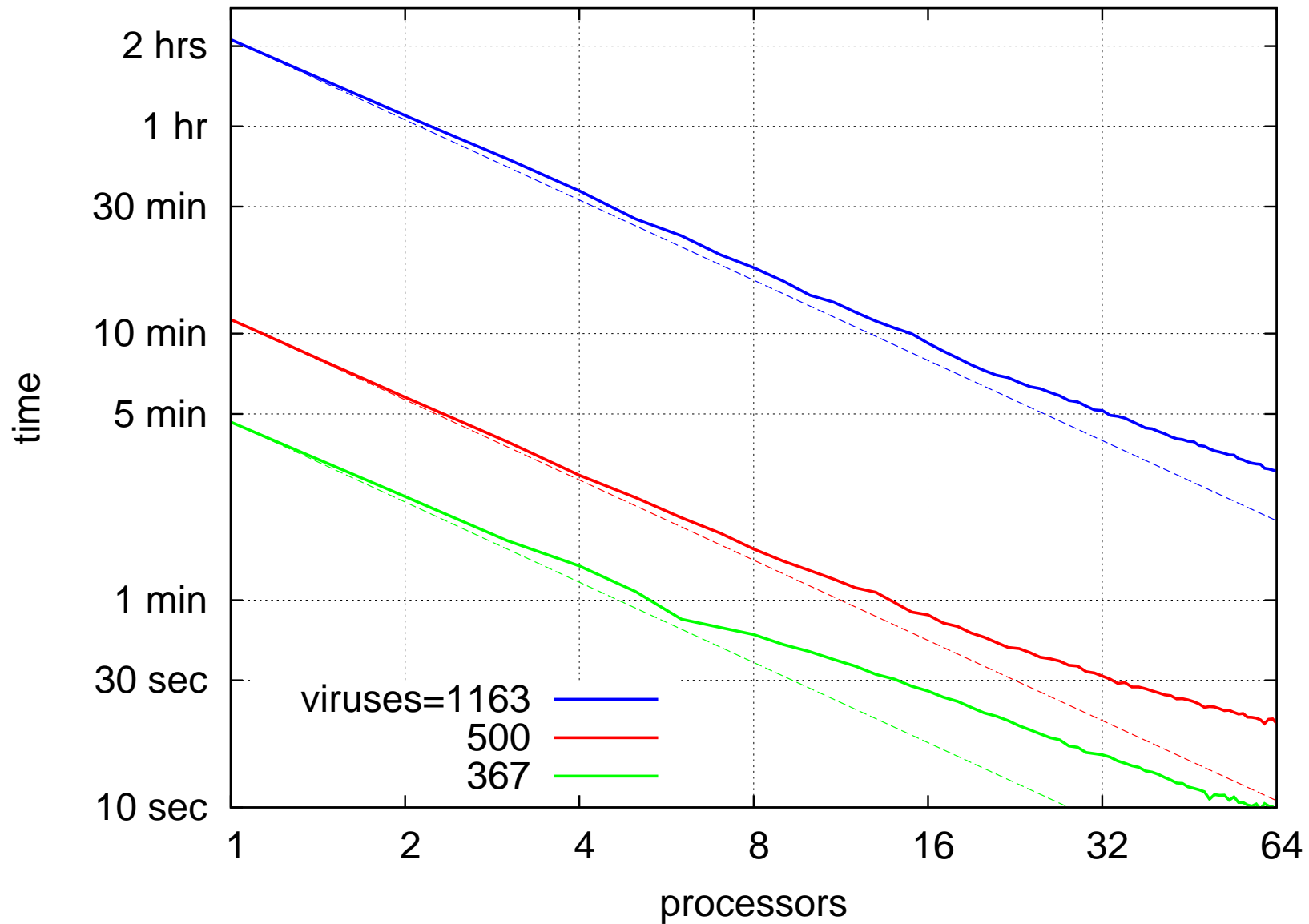
# Comparative Timings



# Comparative Timings



# Timings (restructured code)



Alignment  
Comparison  
Model  
Reassortment  
Network

Shared Multilevel  
Parallel Queues

Motivation

Motivating Example

Performance Bounds  
Generalization  
Extensions

# Shared Multilevel Parallel Queues

# Motivation

- Alignment
- Comparison
- Model
- Reassortment
- Network
- Shared Multilevel  
Parallel Queues
- Motivation**
- Motivating Example
- Performance Bounds
- Generalization
- Extensions

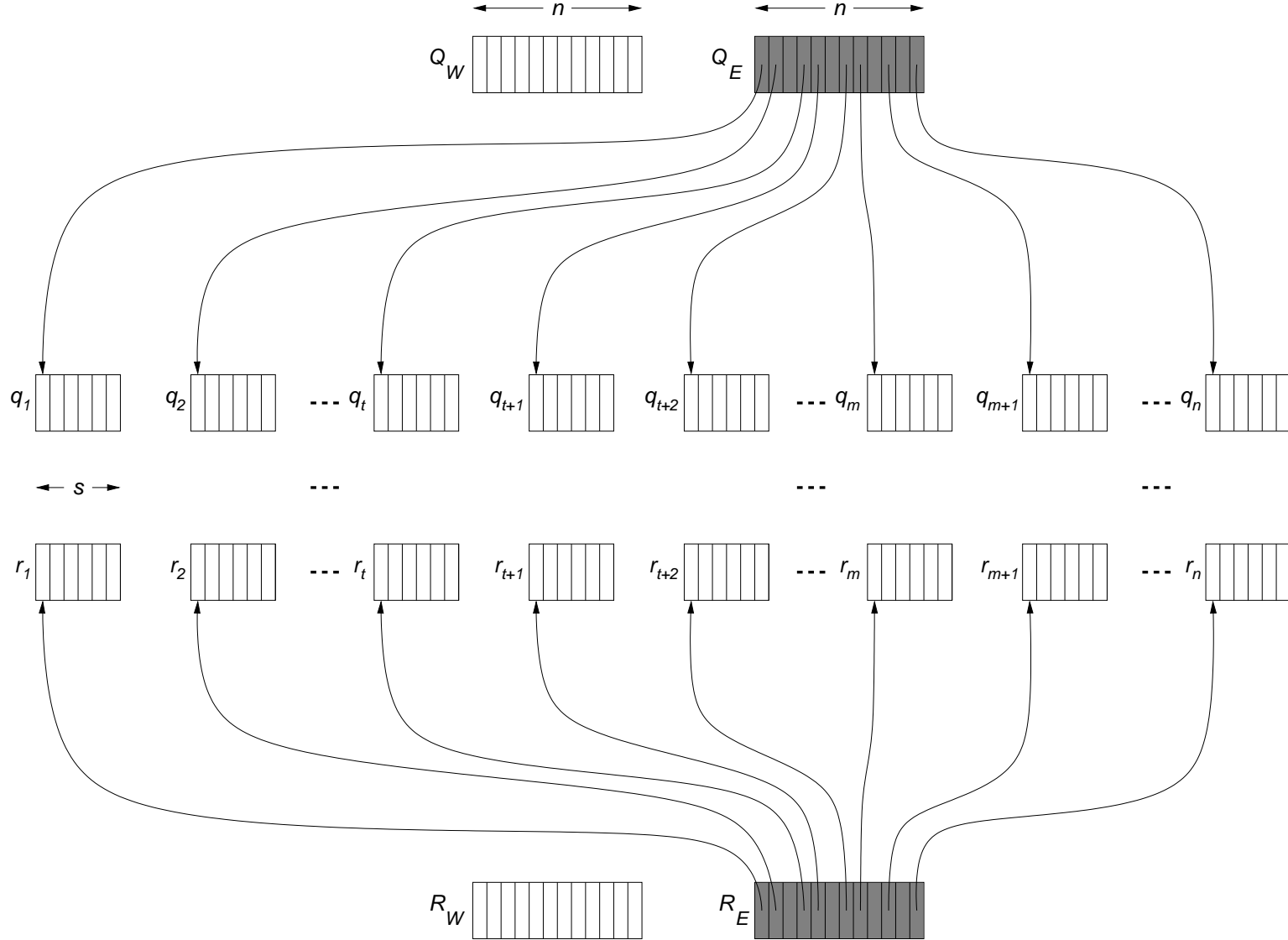
- The use of parallel shared queues is essential for the correct execution of many classes of programs in parallel computers.
- Control of concurrent access to shared queues is possible with `FETCH_AND_ADD` or similar instructions. Unfortunately these instructions generate severe overhead as the number of threads increases.
- A *Shared Multilevel Parallel Queue (SMPQ)* mechanism has been developed that greatly reduces this overhead.

- While the FETCH\_AND\_ADD hardware will permit logically correct operation of shared queues, most parallel computer systems have a limited number of hardware units capable of carrying it out and the performance of queuing operations drops dramatically as the number of active threads increases.
- Consequently, the performance of many algorithms that require frequent use of the FETCH\_AND\_ADD operation is very limited and performance scales very poorly as the number of processors increases.
- At the same time, it is impossible to obtain correct implementations of many important parallel algorithms without using the FETCH\_AND\_ADD operation.

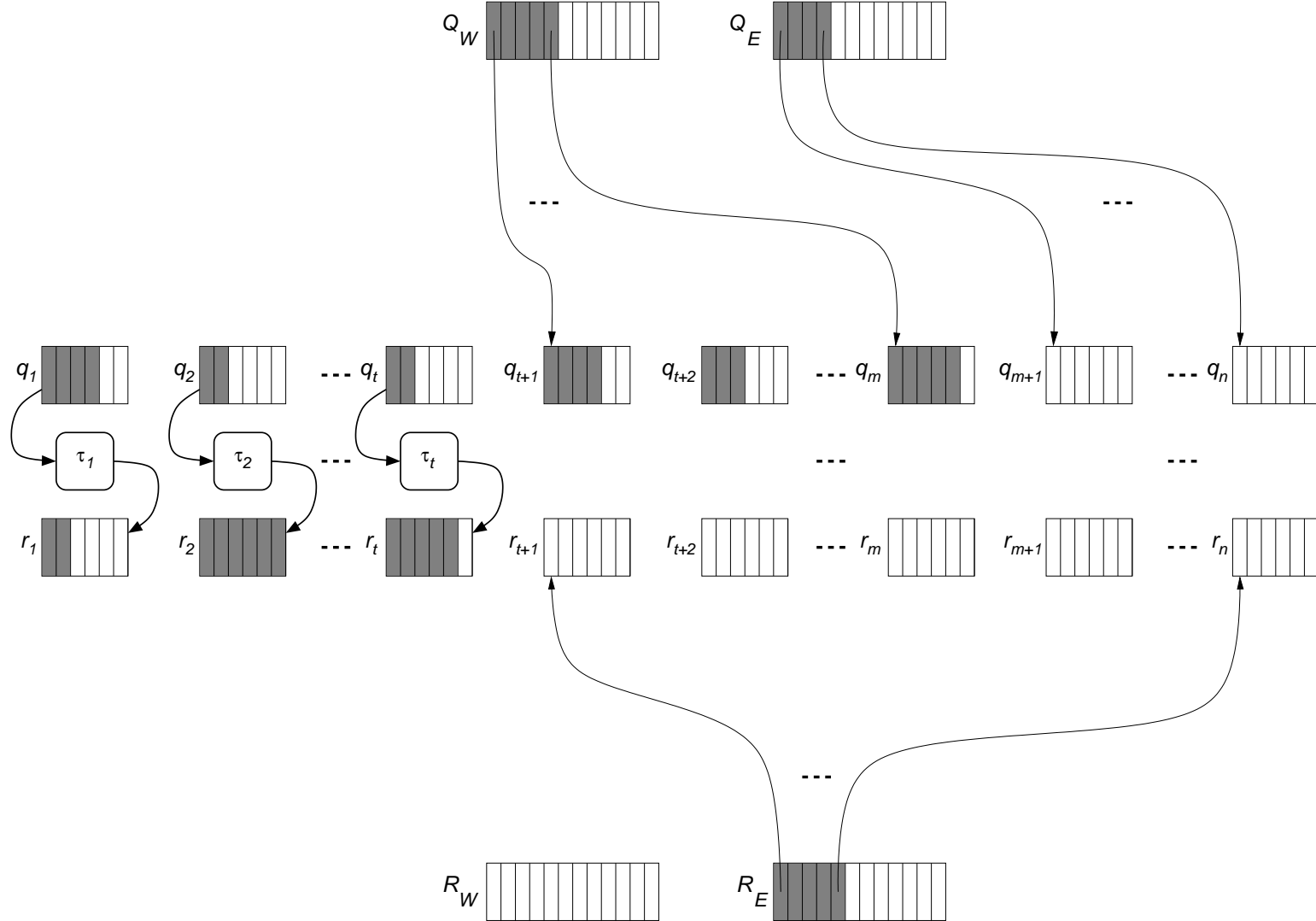
# Motivating Example

- Breadth First Search (BFS) will be used as a motivating example in the exposition of the SMPQ concept.
- BFS is a widely used operation at the heart of many graph algorithms and its efficient parallel implementation shall improve the performance of many existing and proposed parallel graph codes.
- However SMPQs are applicable directly, or with minor modification, to a much wider range of parallel algorithms.

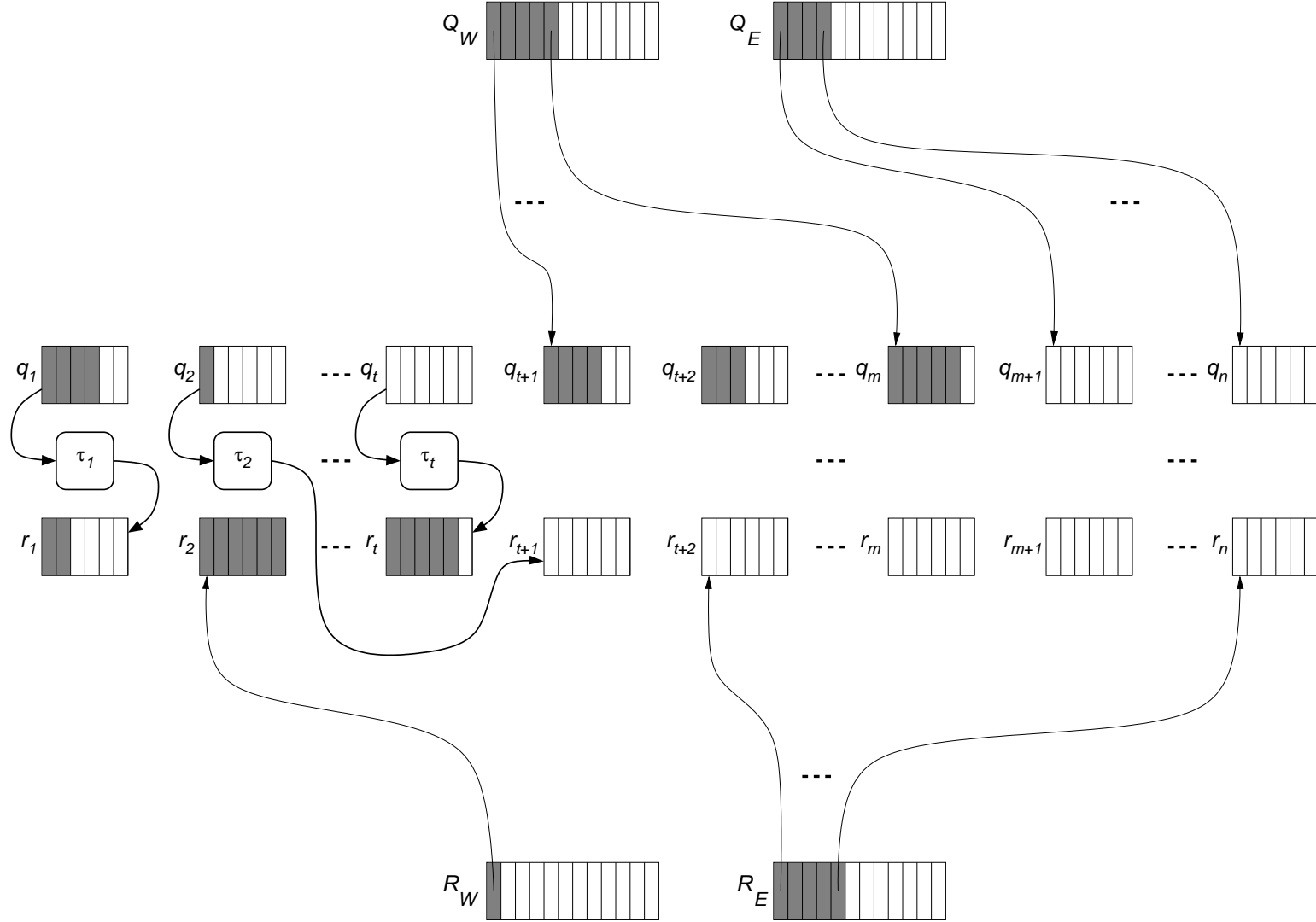




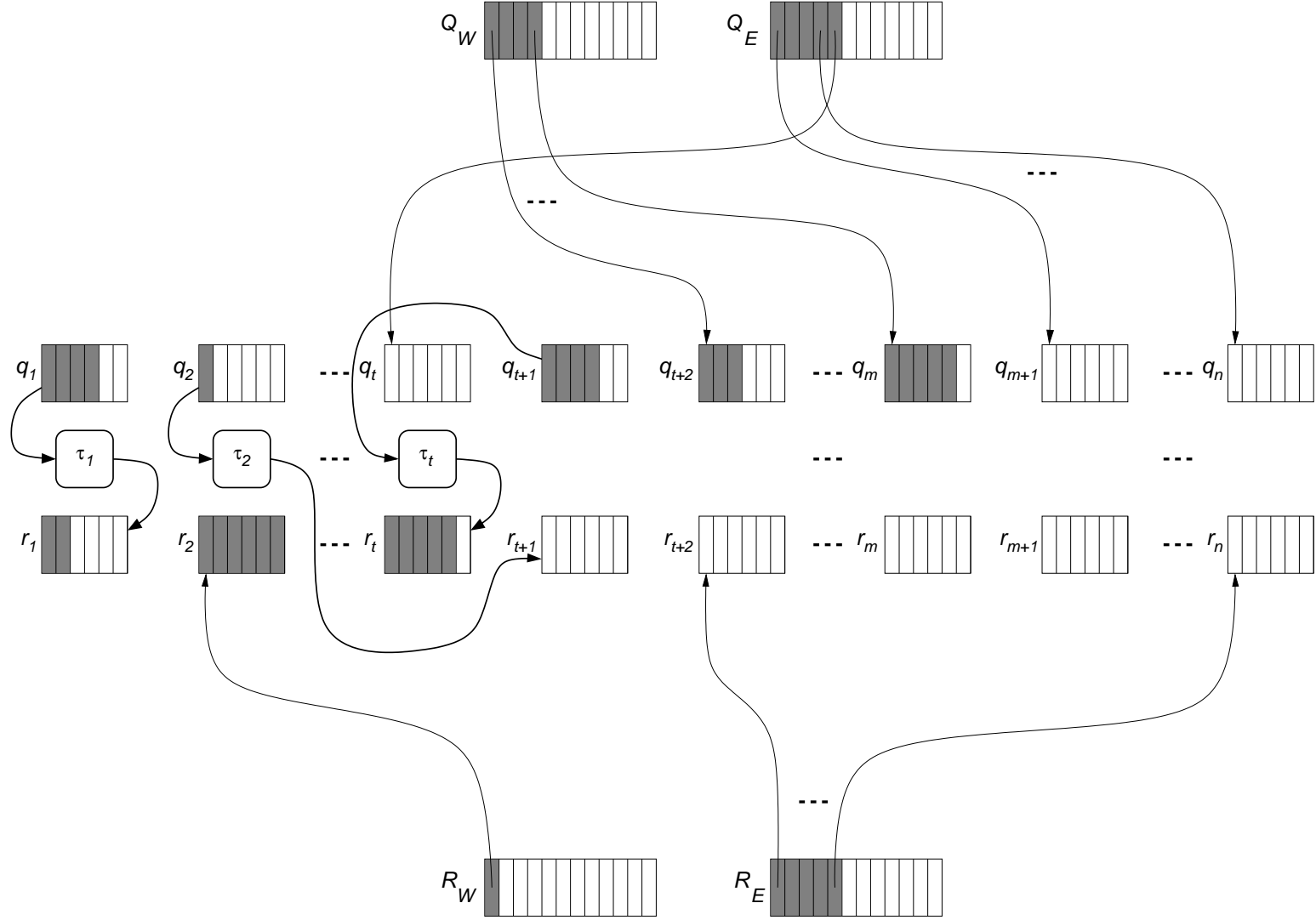
Two first level parallel queues  $Q_W/Q_E$  and  $R_W/R_E$ , each of size  $n$ , hold pointers to serial queues  $q_i$  and  $r_i$  respectively, each of size  $s$ . The total capacity of this system is  $N = sn$ . This figure shows an empty system, thus the pointers are in the queues of empty queues  $Q_E$  and  $R_E$ .

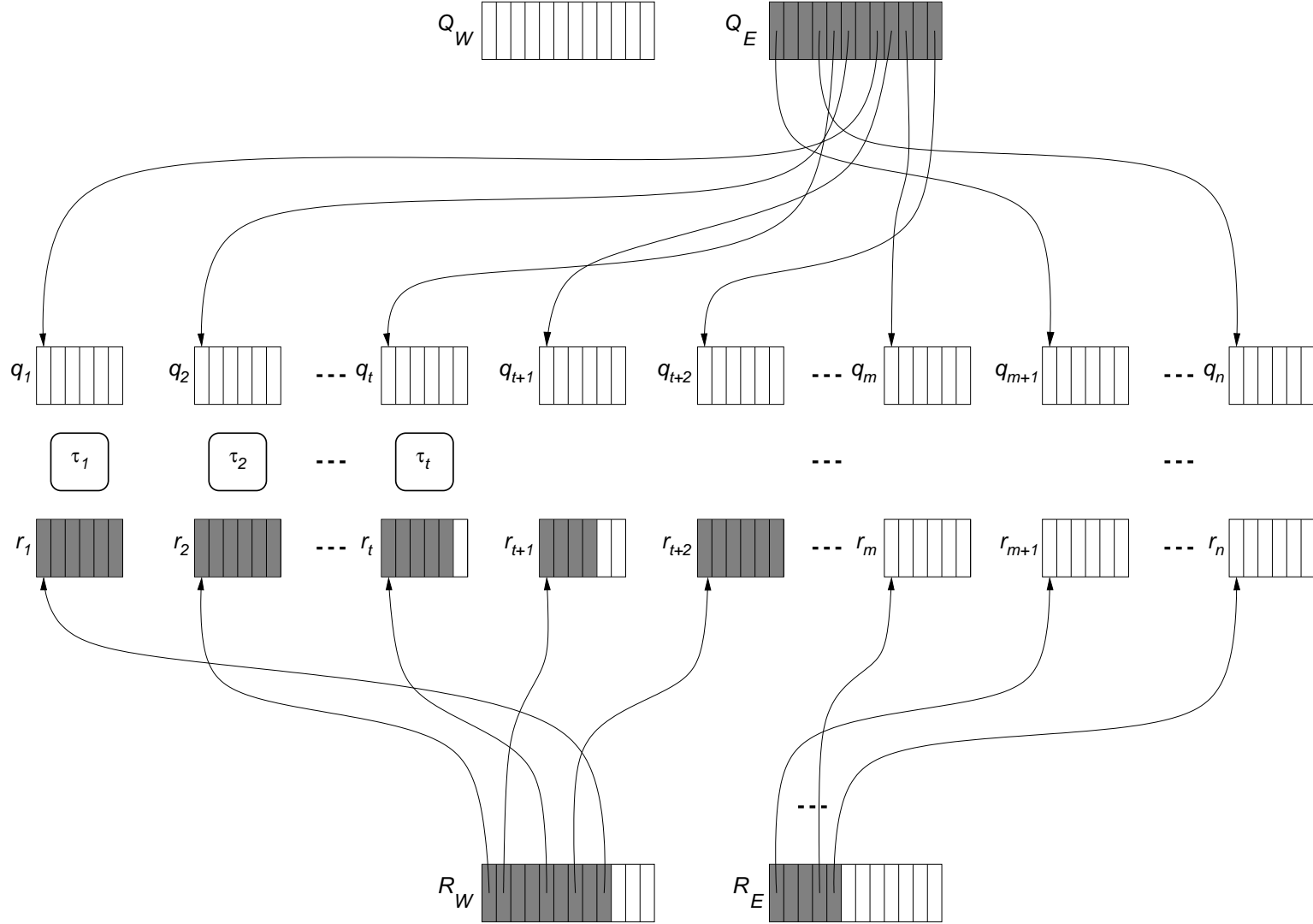


An intermediate stage in breadth first search (BFS). Each thread  $\tau_i$ ,  $0 \leq i \leq t$  has acquired private input(output) queues  $q_i(r_i)$  from  $Q_W(R_E)$ . A thread takes nodes from its private  $q$  and puts nodes on its private  $r$ .  $Q_W$  contains pointers to the remaining  $qs$  that are non-empty.  $Q_E(R_E)$  contains pointers to the remaining  $qs(rs)$  that are empty. Thread  $\tau_2$  has just filled up its output



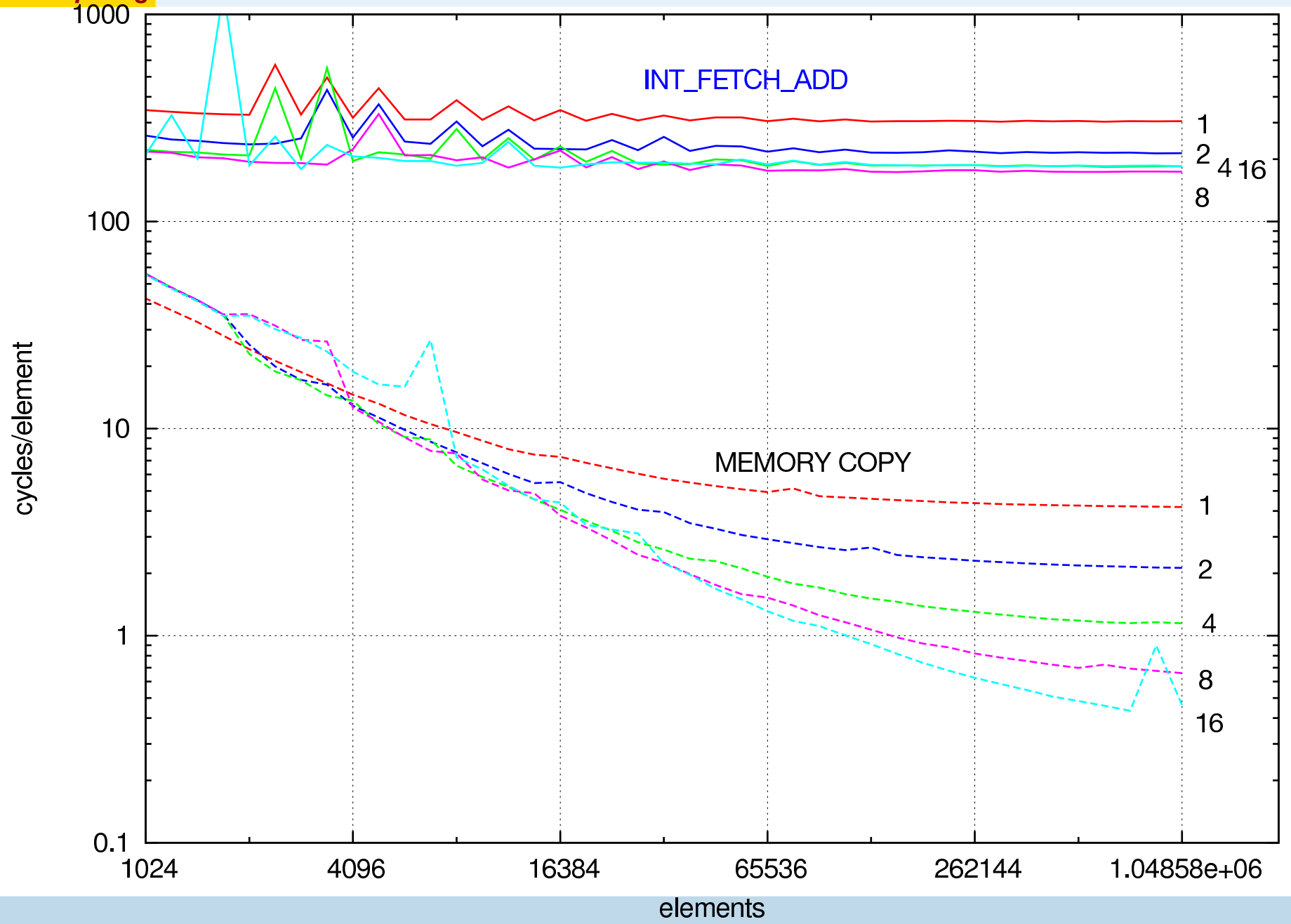
Thread  $\tau_2$  puts its filled queue  $r_2$  in  $R_W$  and gets a fresh output queue  $r_{t+1}$  from  $R_E$ . At this point thread  $\tau_t$  has emptied out its input queue  $q_t$ .



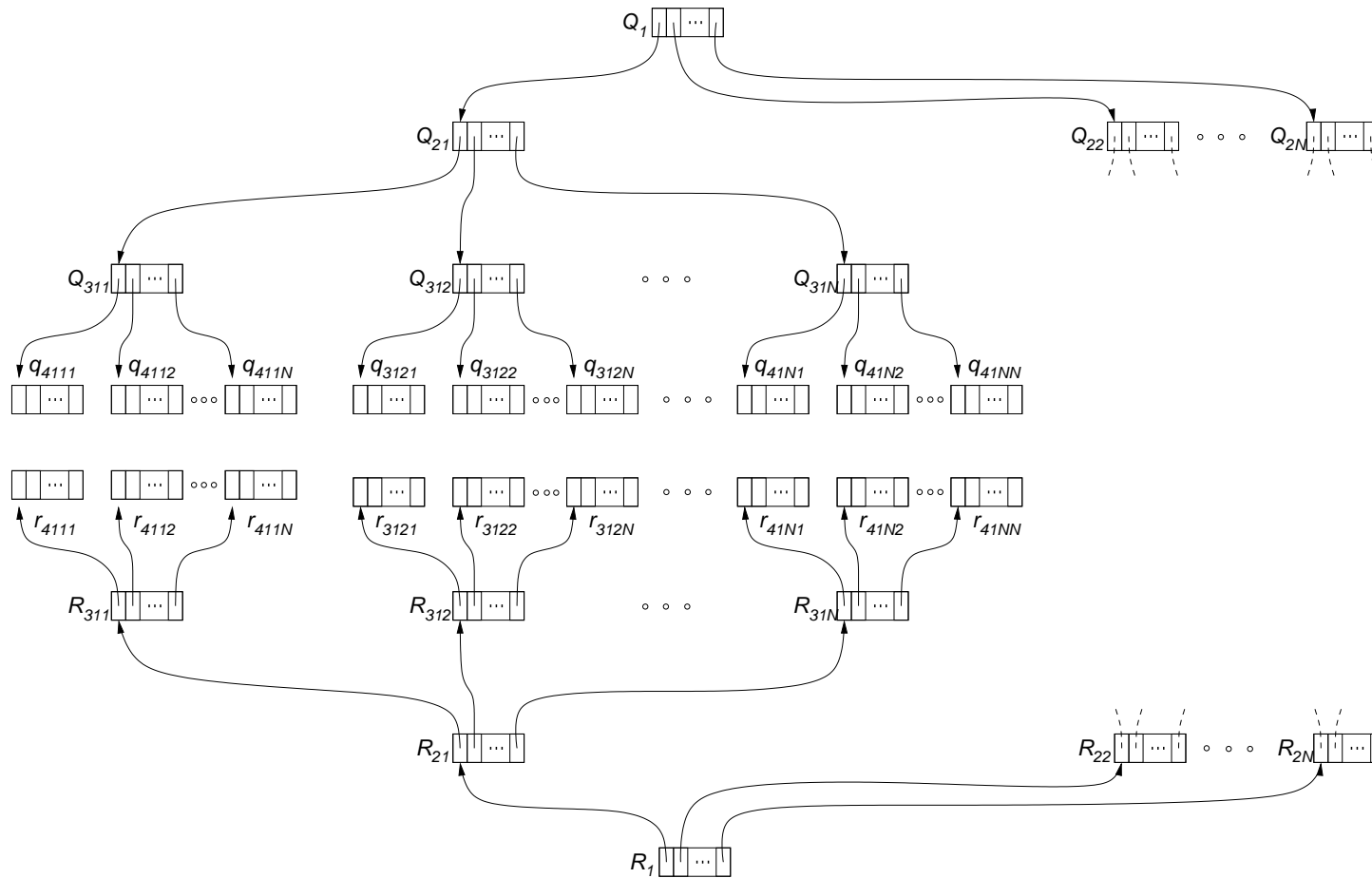


One stage of the BFS is completed, all input queues  $q_i$  (originally in  $Q_W$ ) have been emptied and put in  $Q_E$ . All output queues have been filled (as much as possible) and put in  $R_W$ . At this point the roles of  $Q_W/Q_E$  and  $R_W/R_E$  (as well as  $q_i$ s and  $r_i$ s) are interchanged and the next stage proceeds (the system returns to the configuration shown in first Fig.).

# Performance Bounds



# Generalization



A shared multilevel parallel queuing system with  $k = 4$ . The  $E$  and  $W$  queues have been lumped together for simplicity.

# Extensions

- The shared multilevel parallel queuing mechanism described above has been motivated and presented in the context of Breadth First Search (BFS), where the required capacity of the queue is known *a priori* and is bounded by the total number of nodes in the graph.
- It is equally applicable to other computational problems where a bound on the total queue size is known. However, there may be cases where a bound on the total queue size is not known beforehand.
- In such cases it is straightforward to extend the approach so that when the system sees the total capacity approach exhaustion, it allocates additional queues (in dynamic memory) and inserts them appropriately in the data structure. The overhead of carrying out these allocations will be hidden by the activity of the active threads, of which there may be thousands or even millions in a massively multithreaded system.