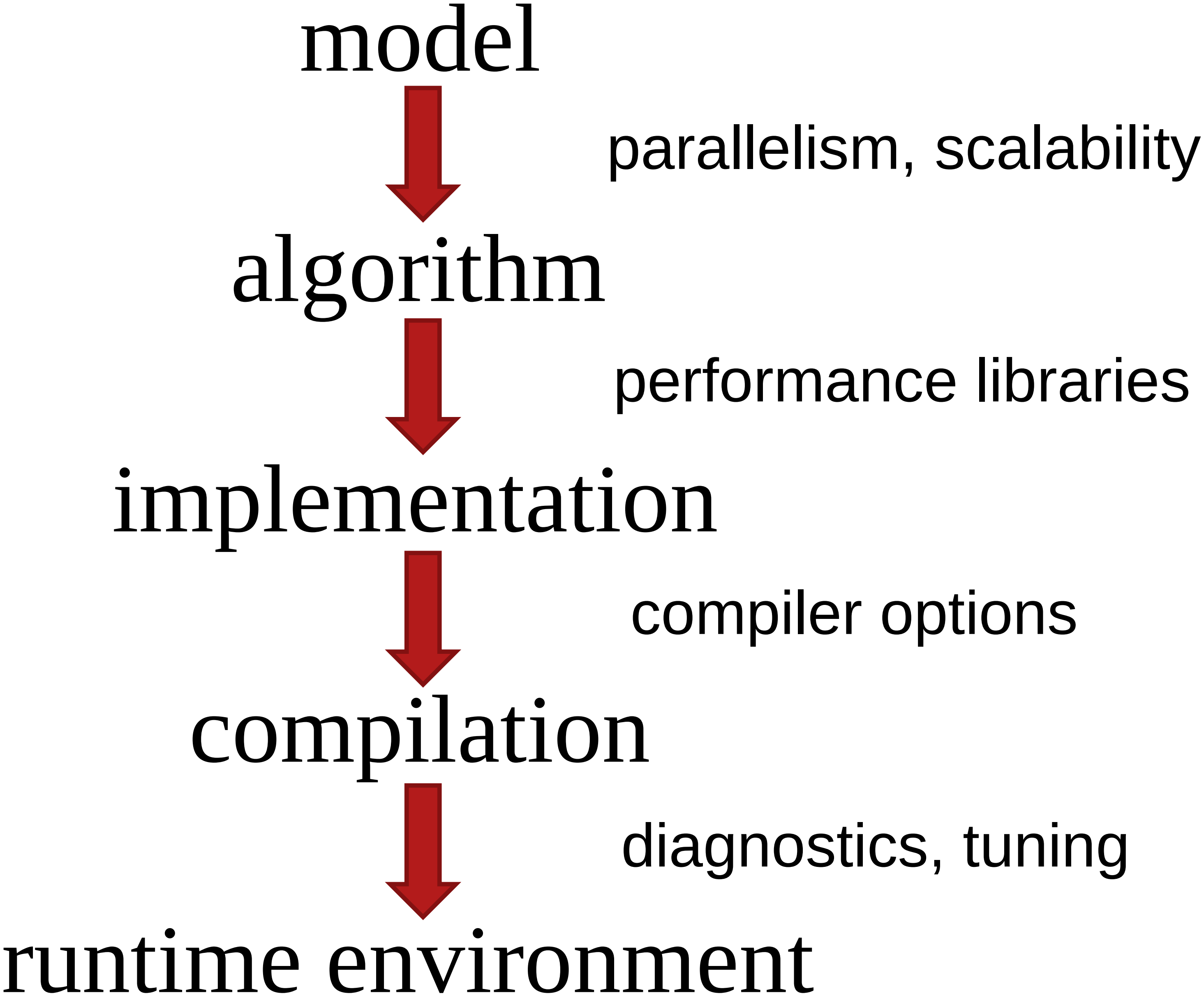
Optimization and Scalability

on Ranger and Lonestar

Drew Dolgert



model

parallelism, scalability

algorithm

performance libraries

implementation

compiler options

compilation

diagnostics, tuning

runtime environment

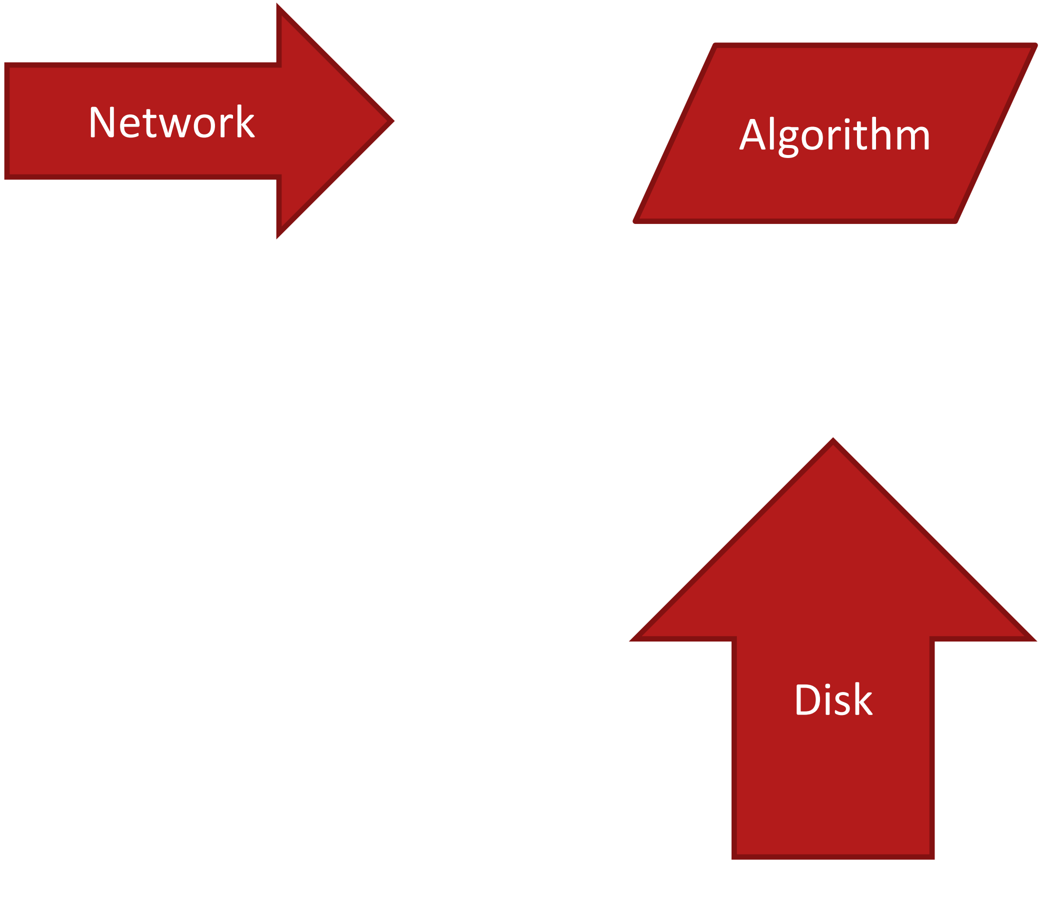
Libraries

Performance	Math Libs	Method Libs	Applications	I/O
gprof	fftw	petsc	Amber	netcdf
tau	GotoBLAS	scalapack	NAMD	hdf5
papi	Metis/parmetis		charm++	
	MKL 10.0		Gamess	
	Gnu Scientific Library			

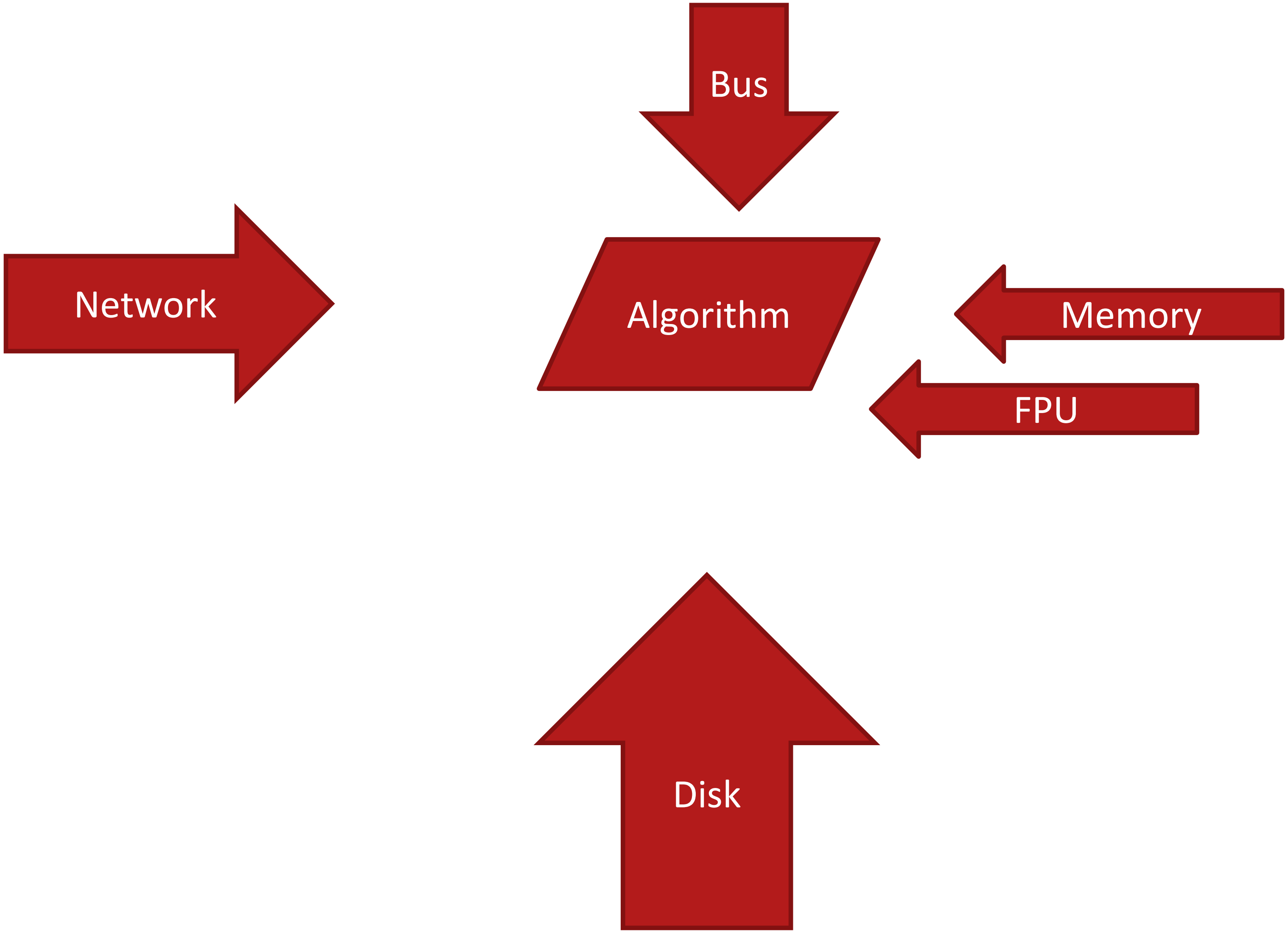
Exercise Libraries

2.2 Compare libraries and hand-written code

Model of an Algorithm’s Environment



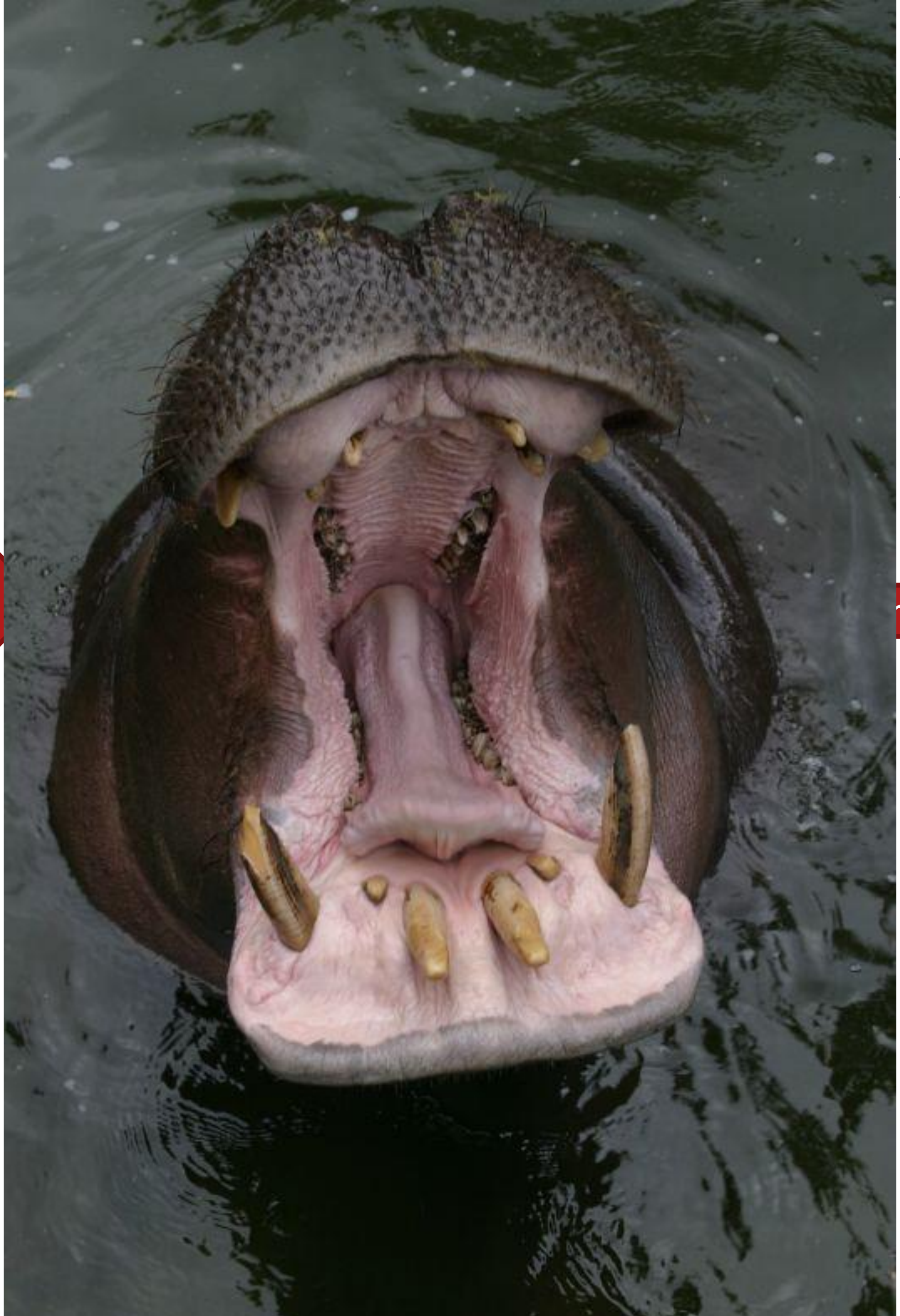
Model of an Algorithm's Environment



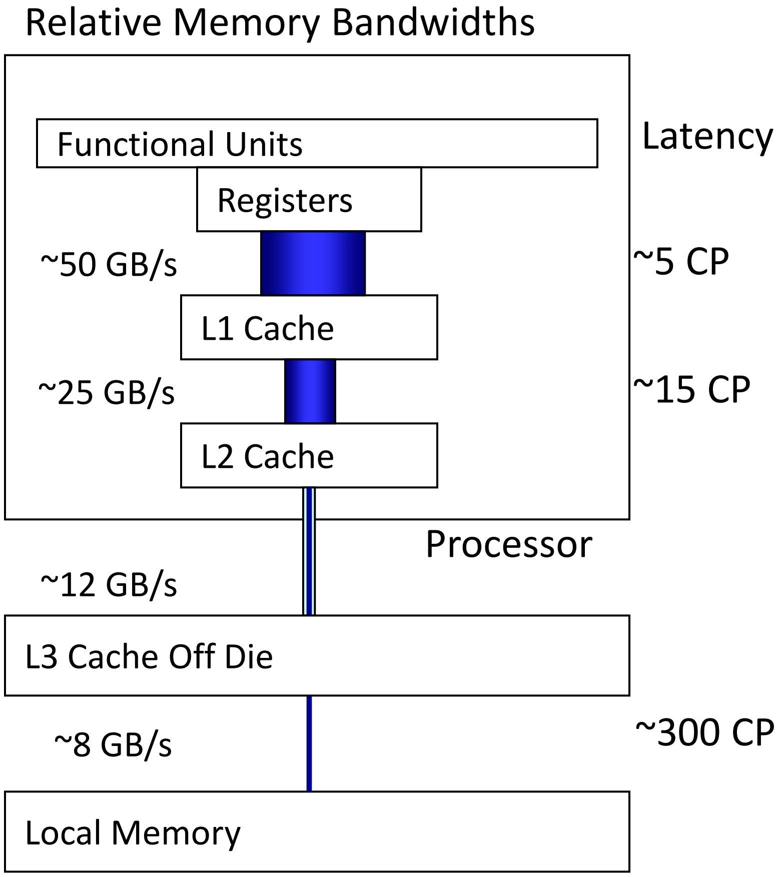
Model ironment

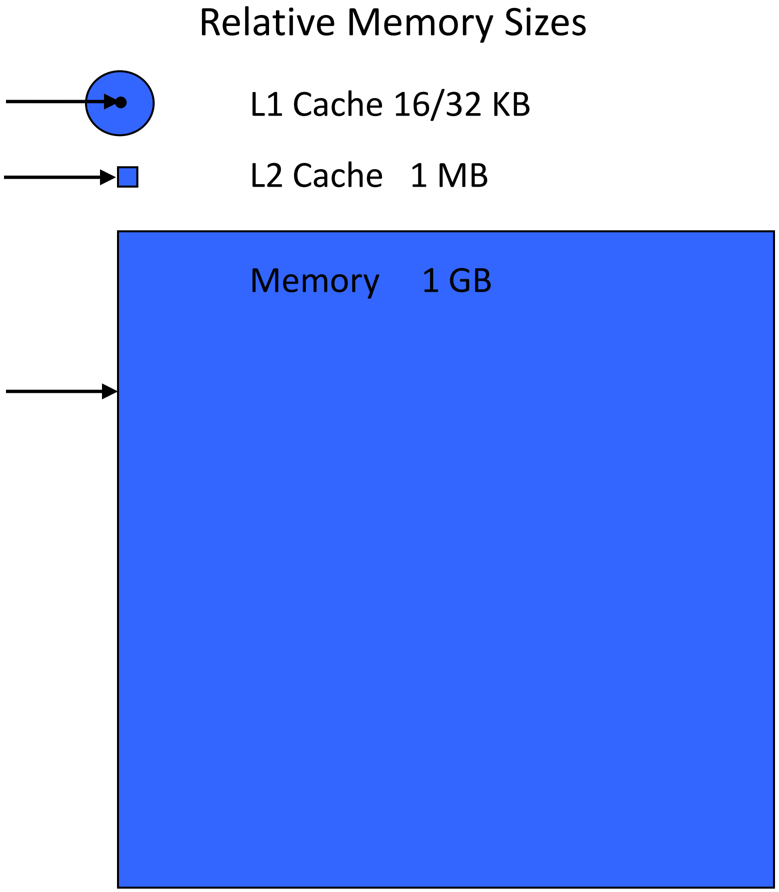
Network

mory

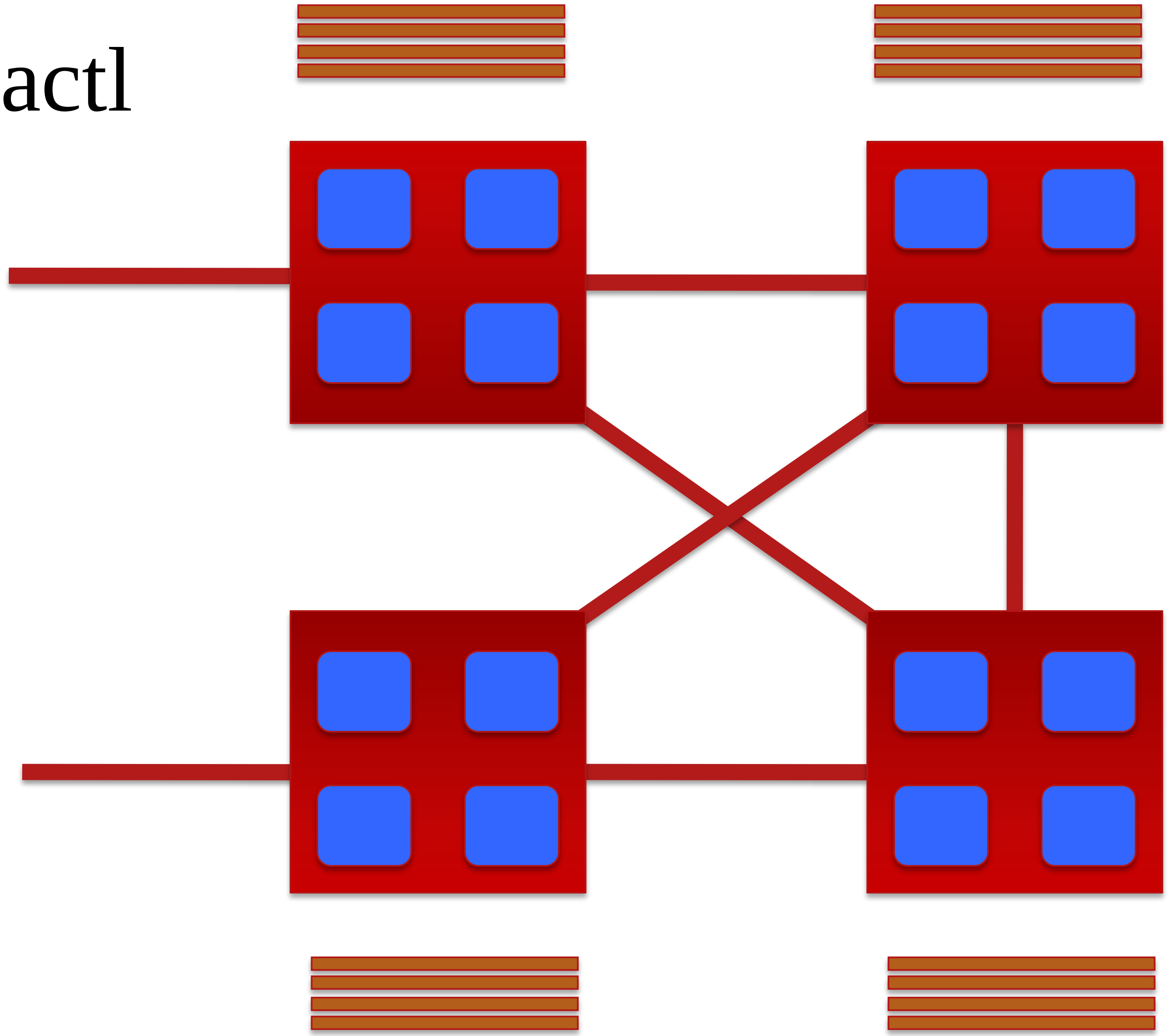


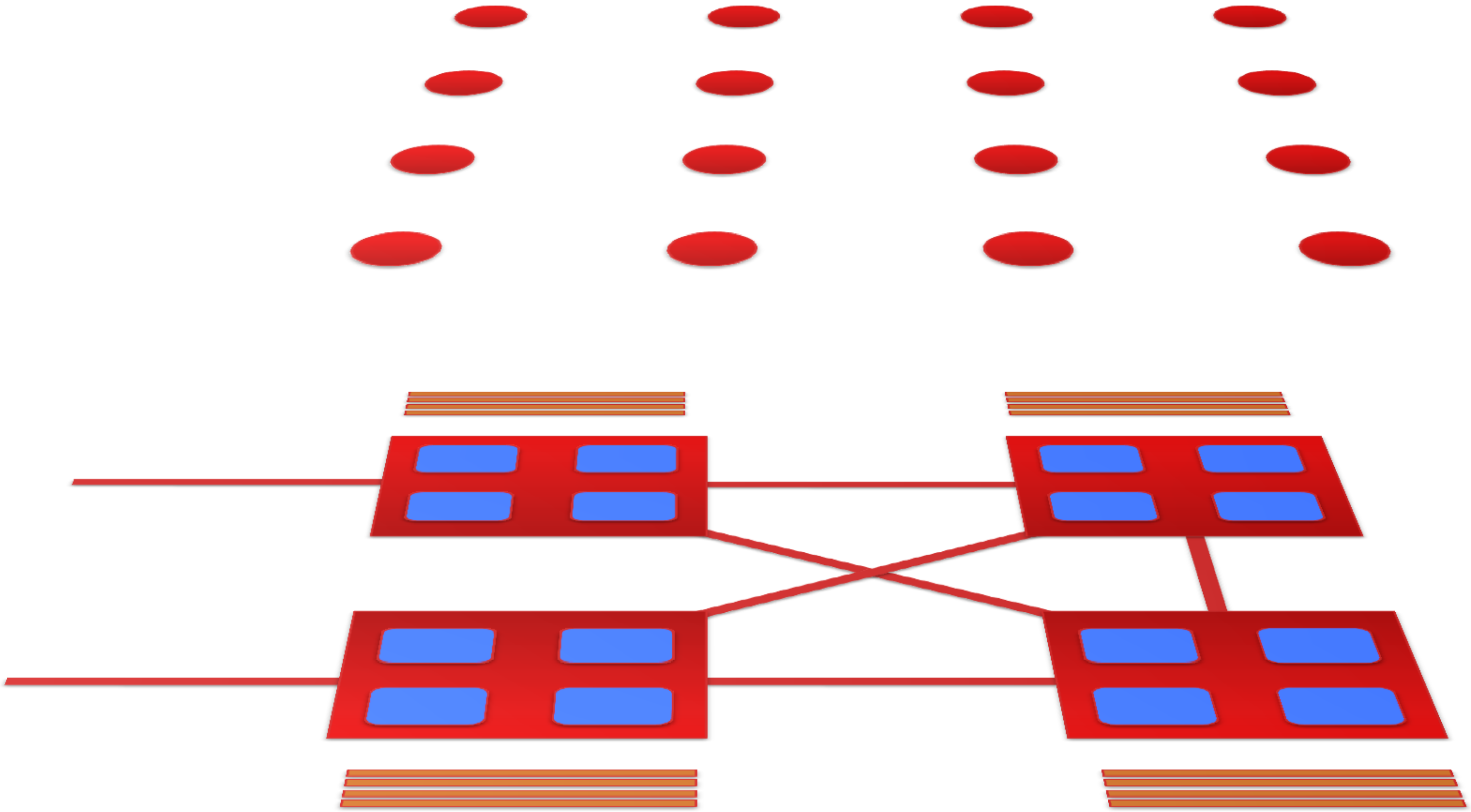
Memory: The Long Pole in the Tent





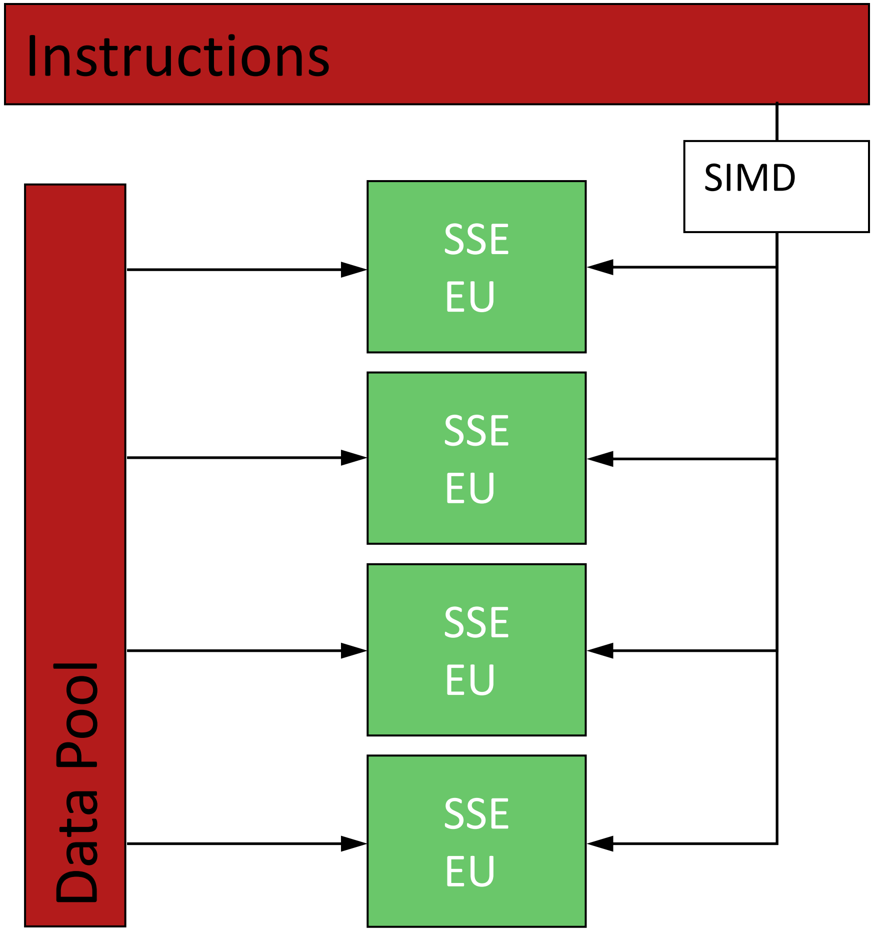
numactl





Use your chip – vectorize.

```
for (i=0;i<N;i++) {
    y[xIdx] = sqrt(psi[i][j]);
}
```



cat /etc/procinfo

Let the compiler roam.

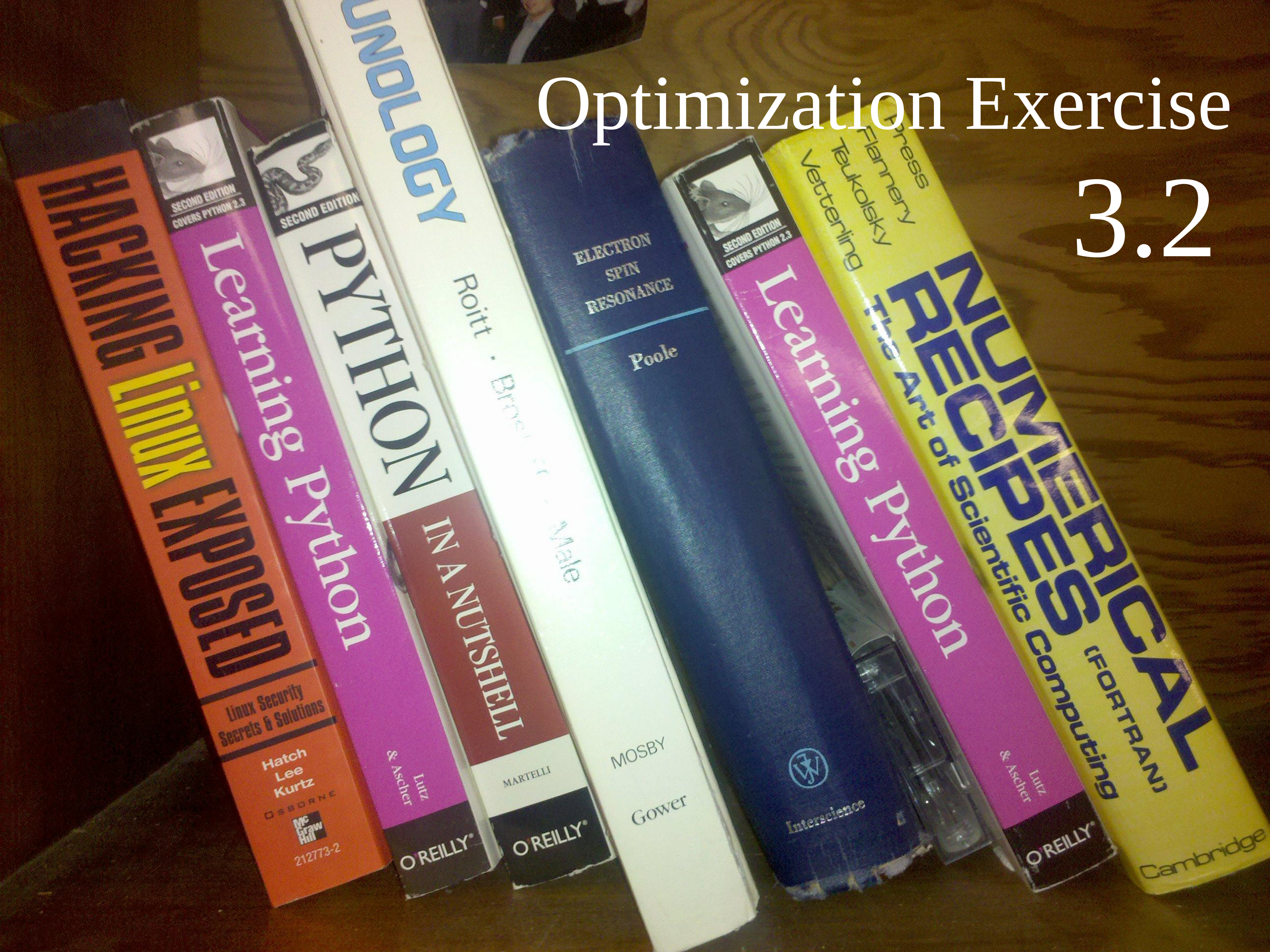
Interprocedural Optimization - -ipo

But watch it.

-g -O1 -O2 -O3

Optimization Exercise

3.2



This code runs well.

It has to run this many times.

SU ≈ cpu-hour

$$\text{time} = \_ + \_ + \_$$

Parallel Random Access Model of the Machine

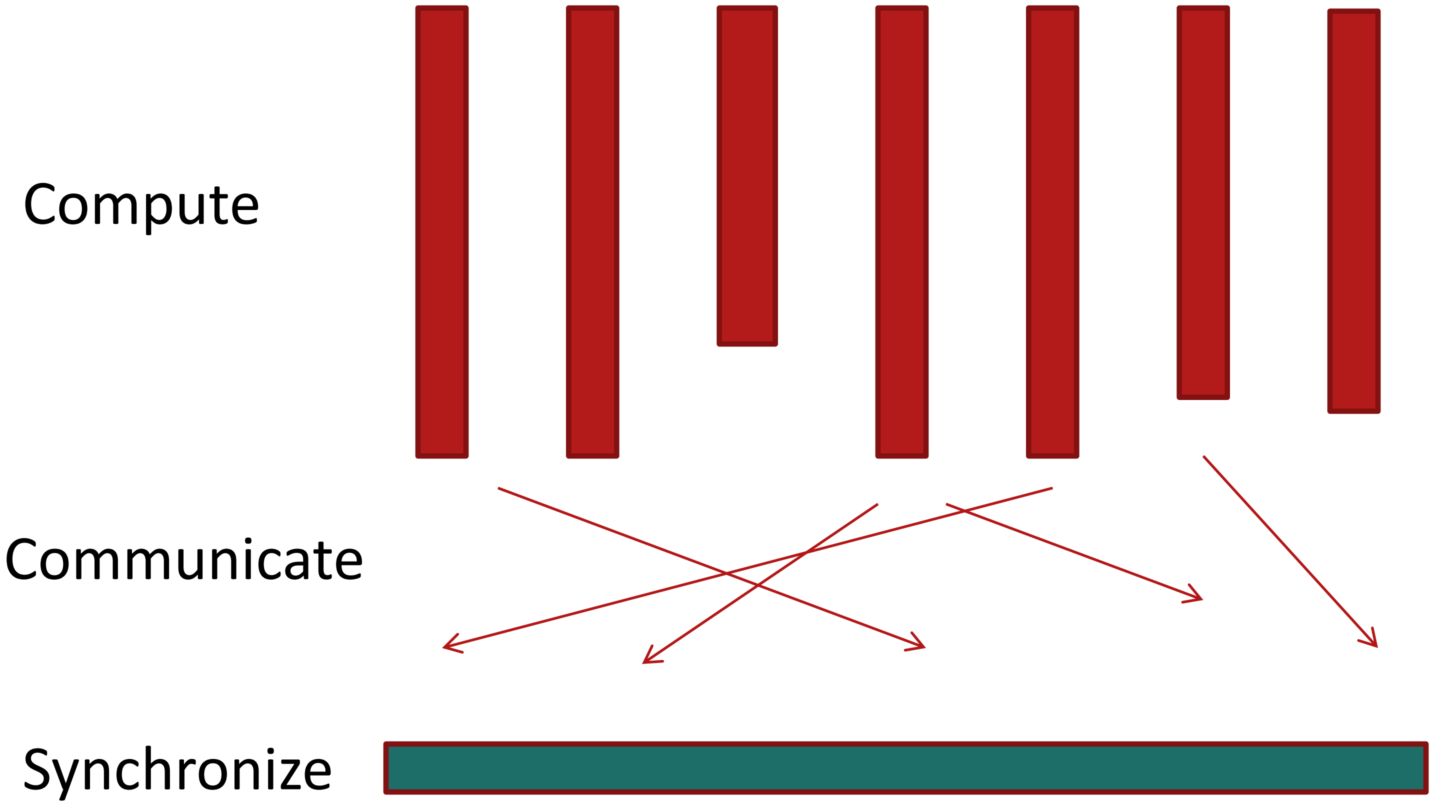
- Multiple processors.
- Single shared memory.
- Every processor accesses memory in unit time.

LogP Model of Machine

- Latency of communication medium
- Overhead of sending and receiving a message
- Gap between two send / receive operations
- Processing units, the number of them.

What year?

Bulk Synchronous Parallel Model for Computation



Compute

Communicate

Synchronize

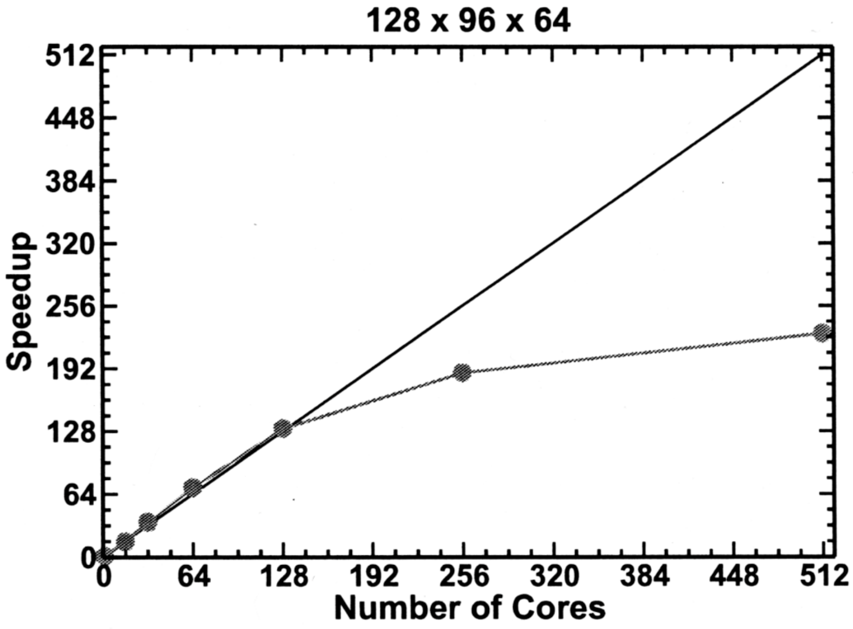
What year?

Exercise

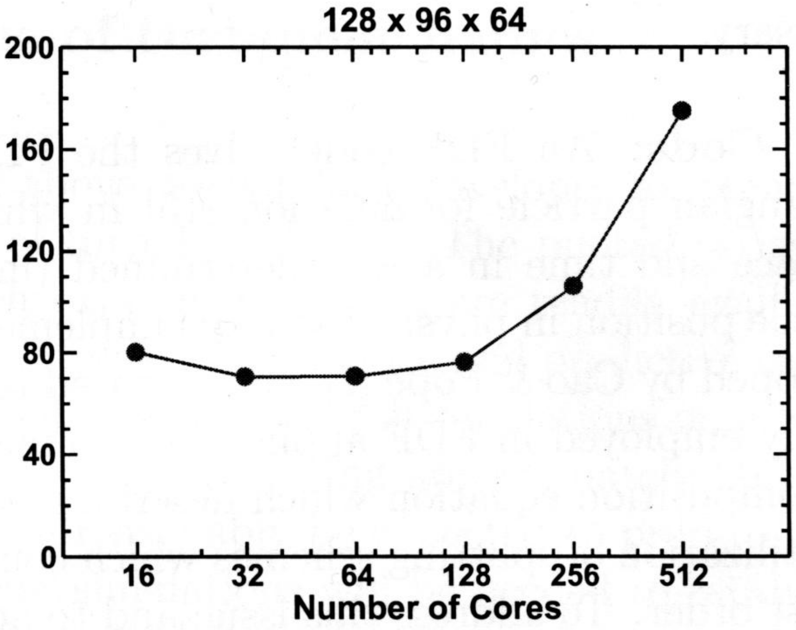
Focus:

- How much work to do.
- When and how much to communicate.
- Structure of communication.

4.1 Analyze a Parallel System

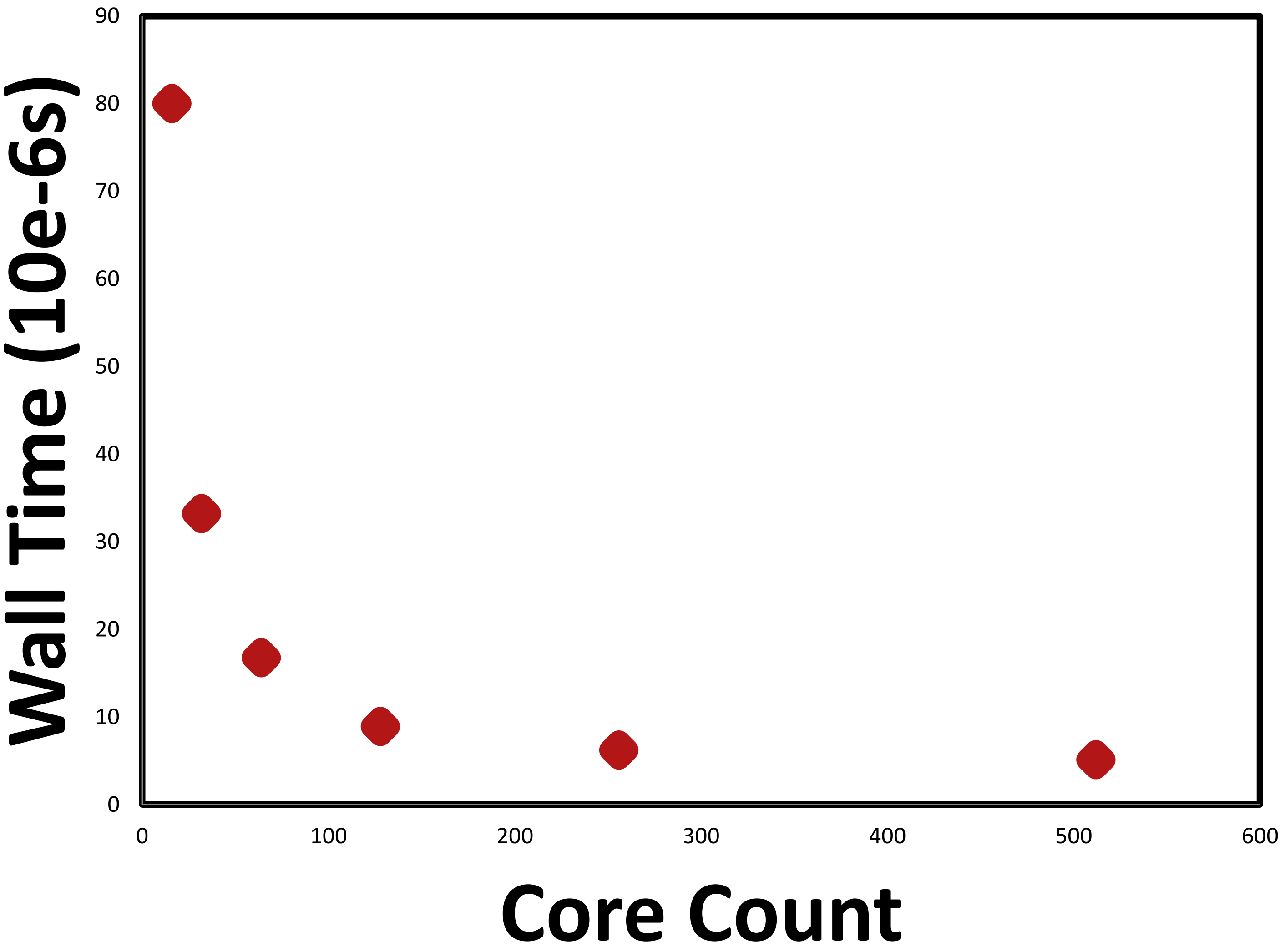


(a)

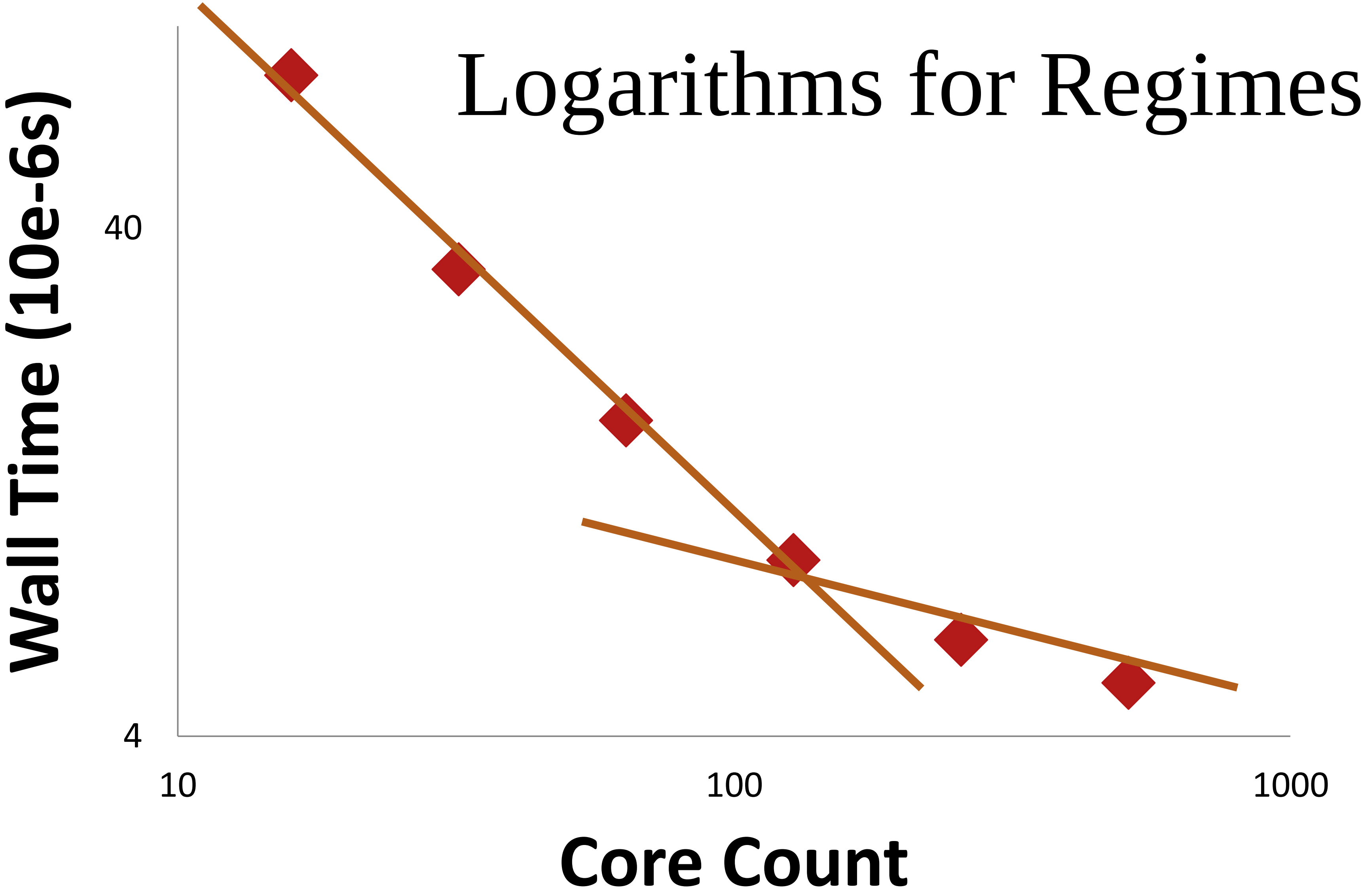


(b)

It is seen that for the chosen grid resolution the LES code exhibits linear scalability up to 128 processors and reasonable scalability up to 256 processors.


Wall Time (10e-6s)
90
80
70
60
50
40
30
20
10
0
0
100
200
300
400
500
600
Core Count

Logarithms for Regimes

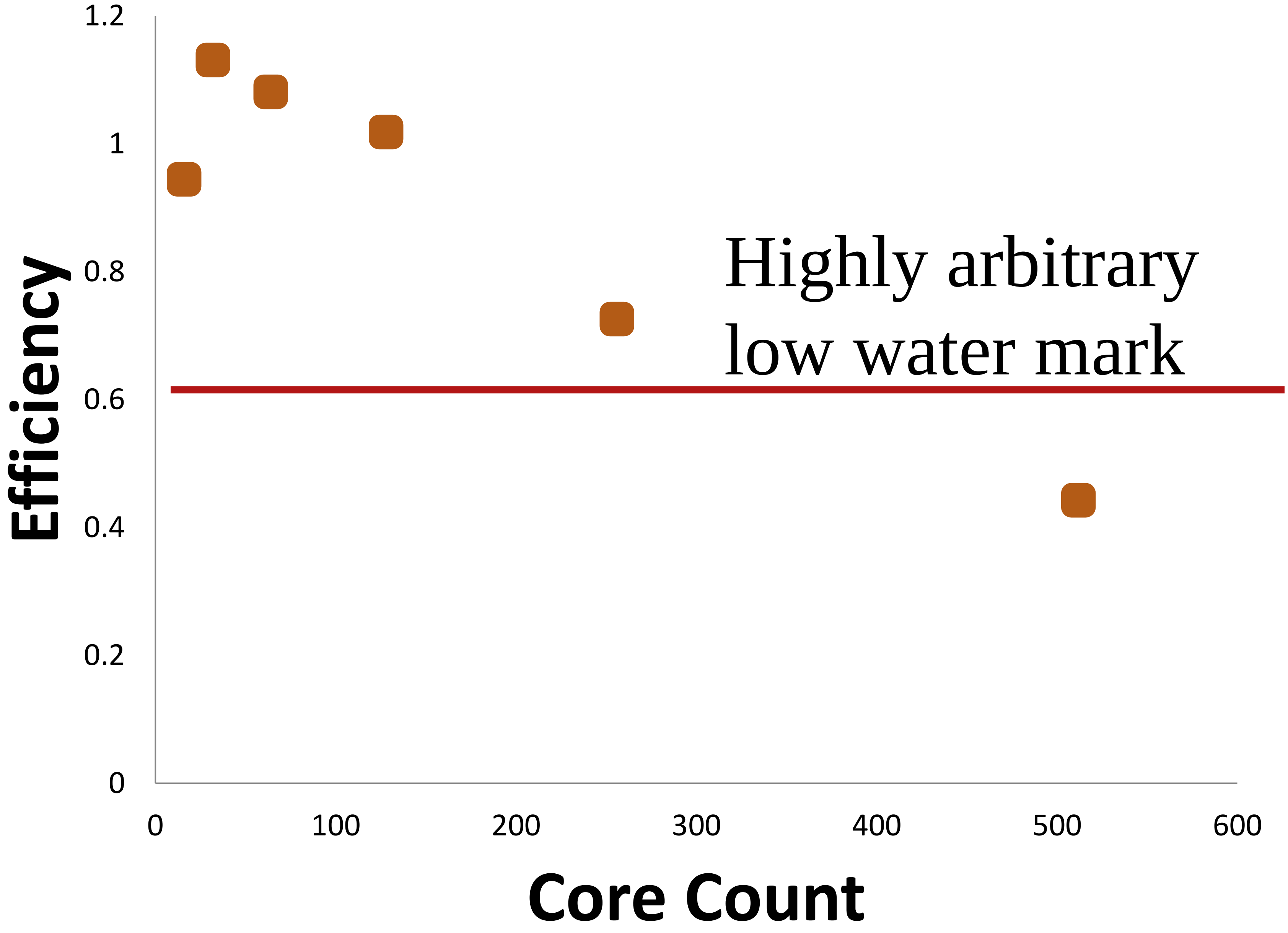


Efficiency

$$\frac{\text{How much an N-way parallel job does}}{\text{How much an N serial jobs do}}$$

Efficiency

$$\frac{\text{Time for 1-way job} / N}{\text{Time for N-way job}}$$

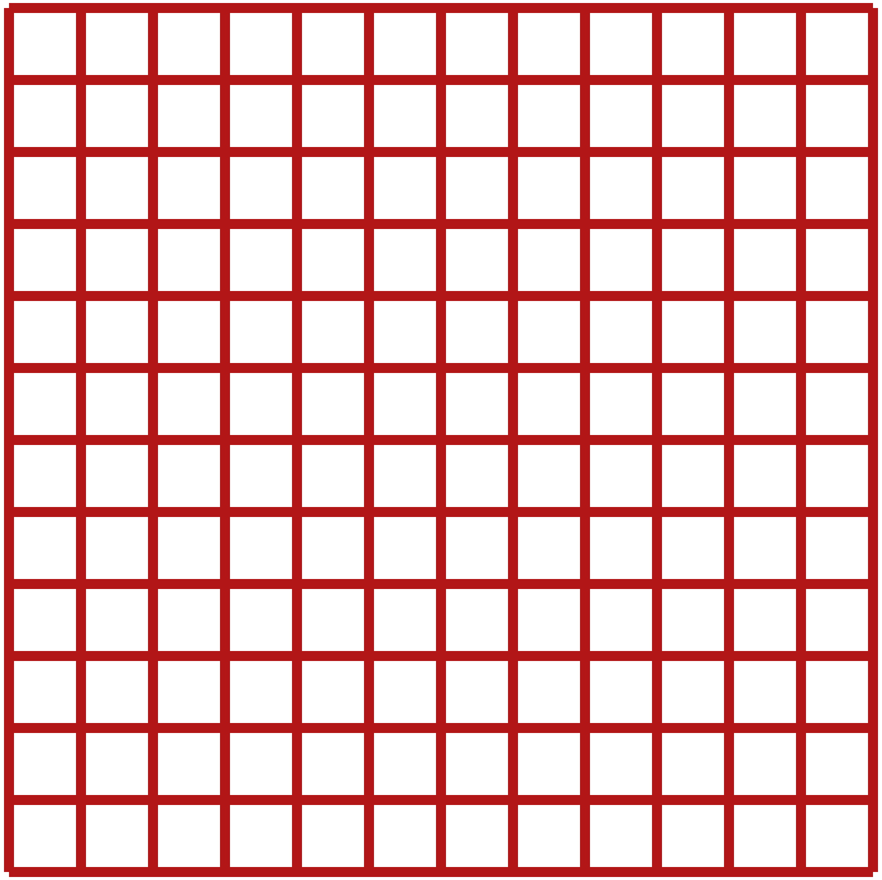


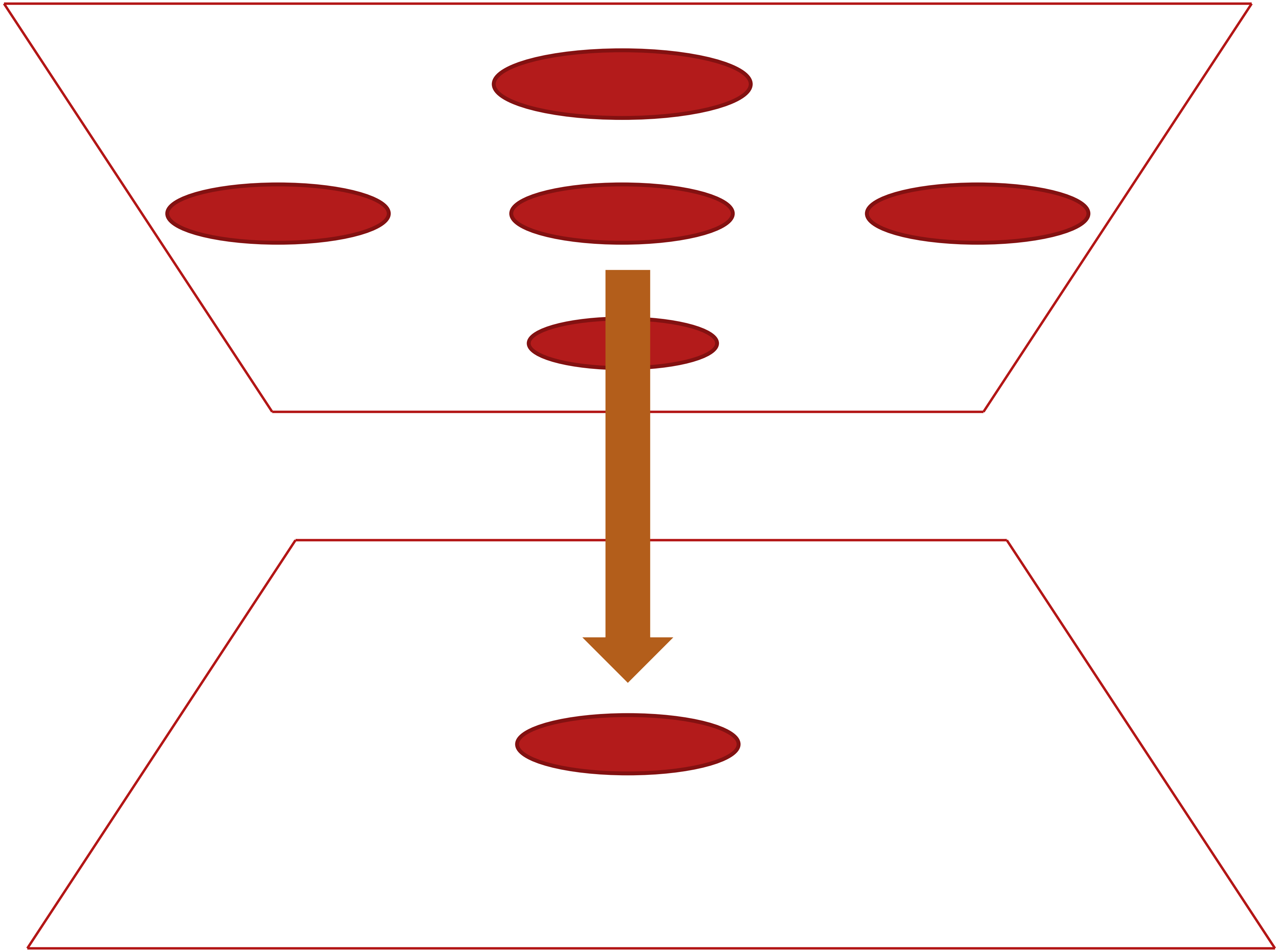
Highly arbitrary low water mark

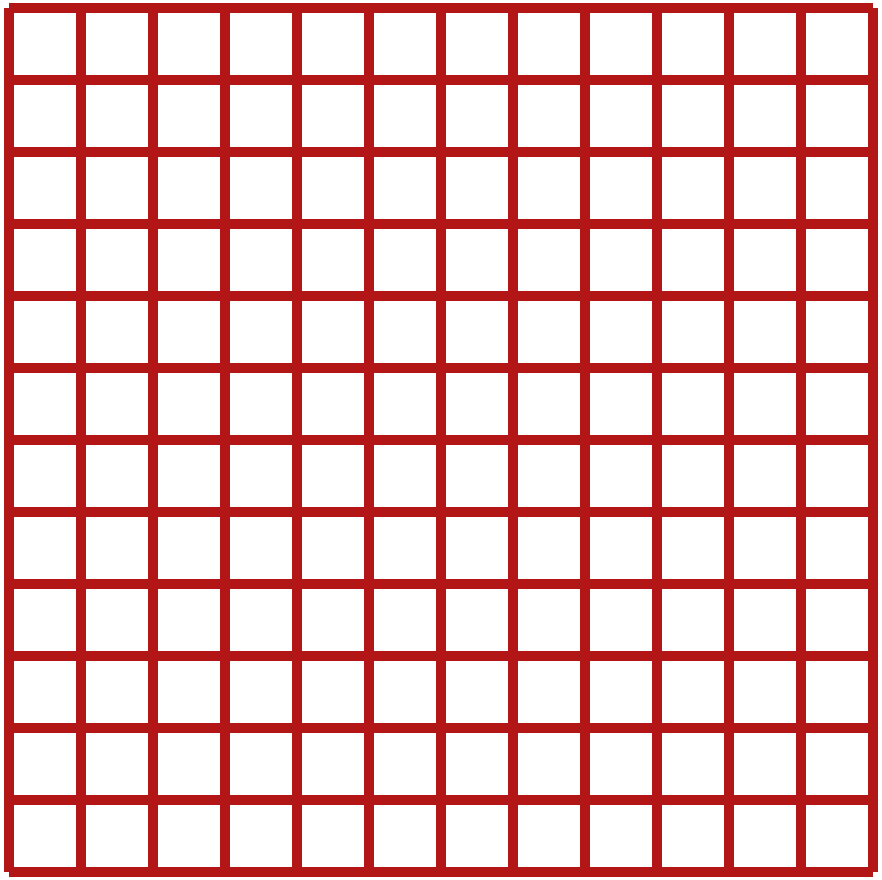
1.4 Allocation MadLib

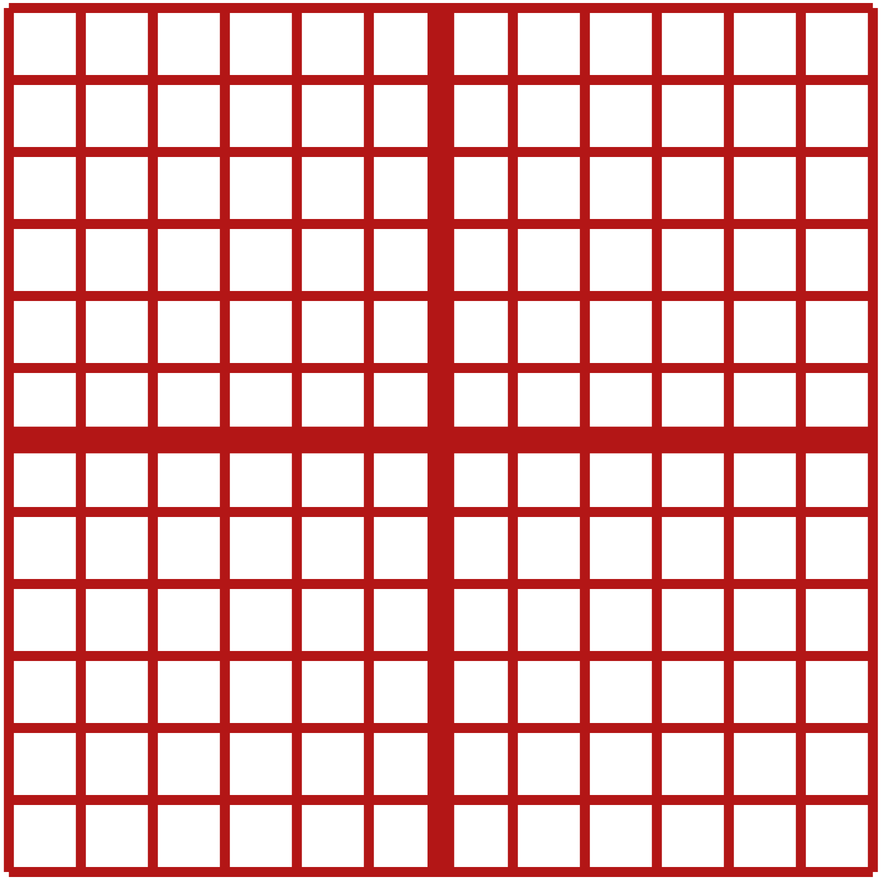
We expect improvements in scalability with increasing problem size.

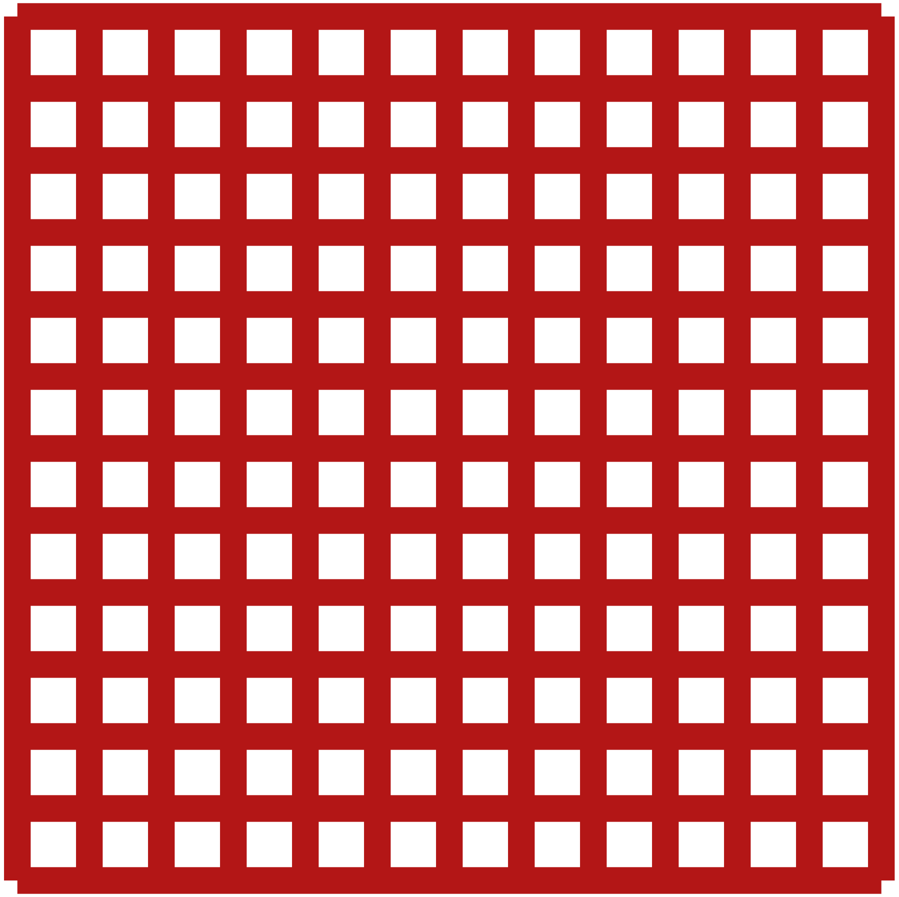
Communication *versus* Computation

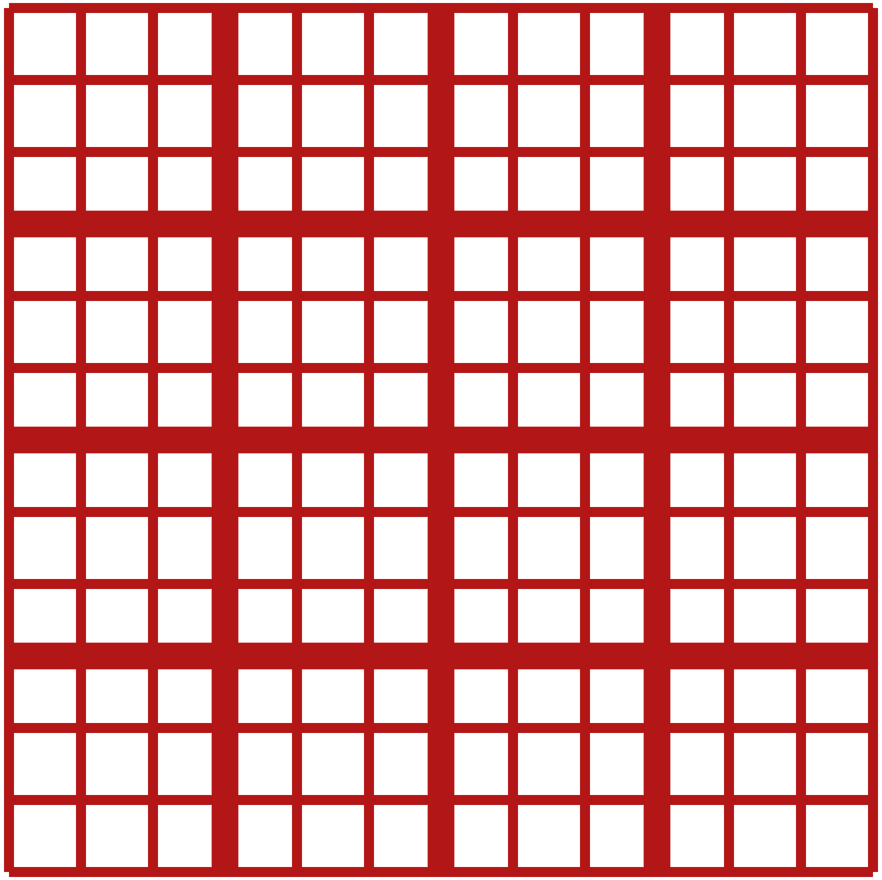








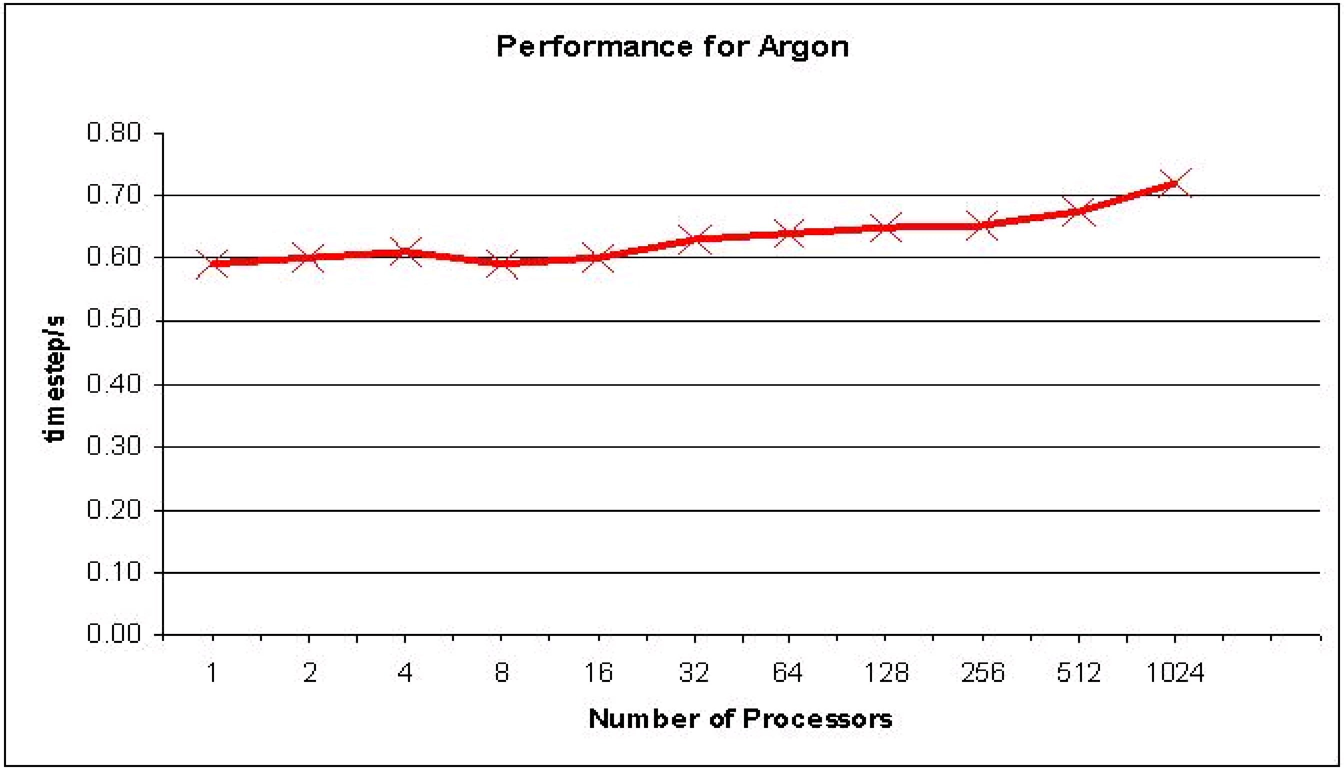


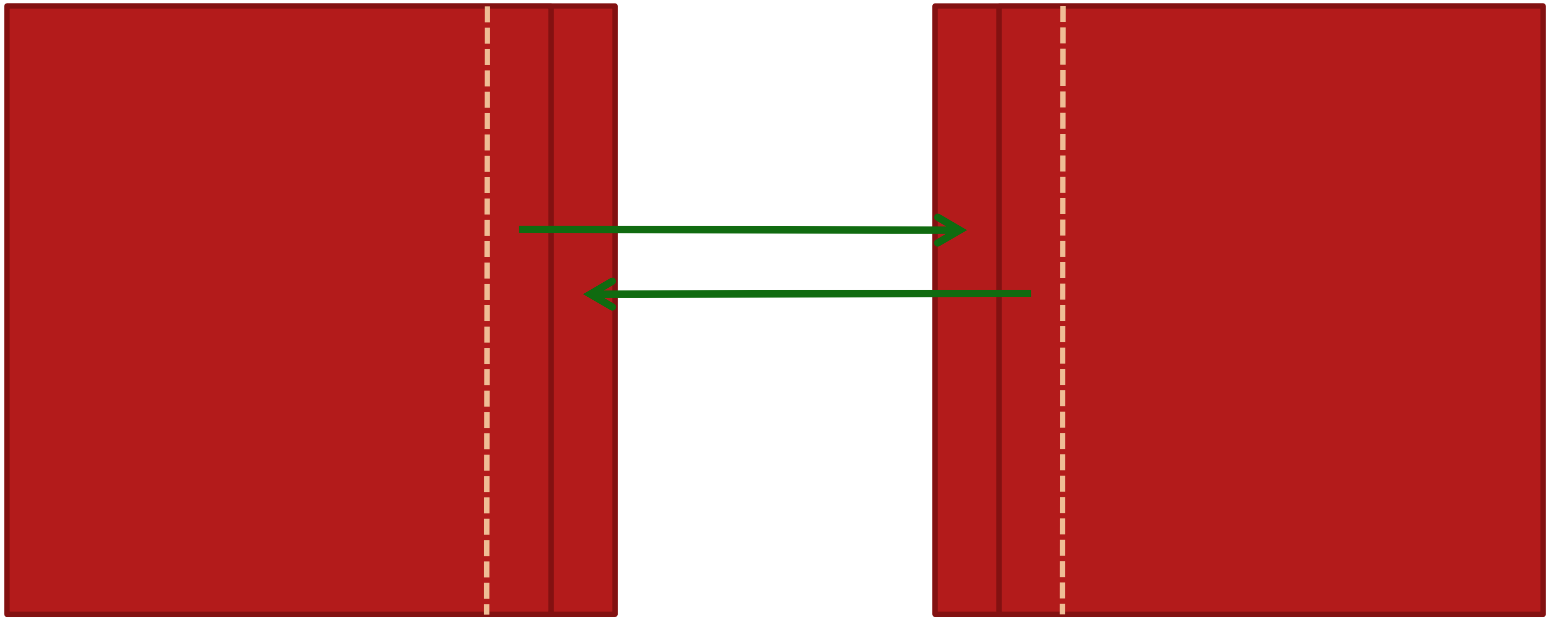


W

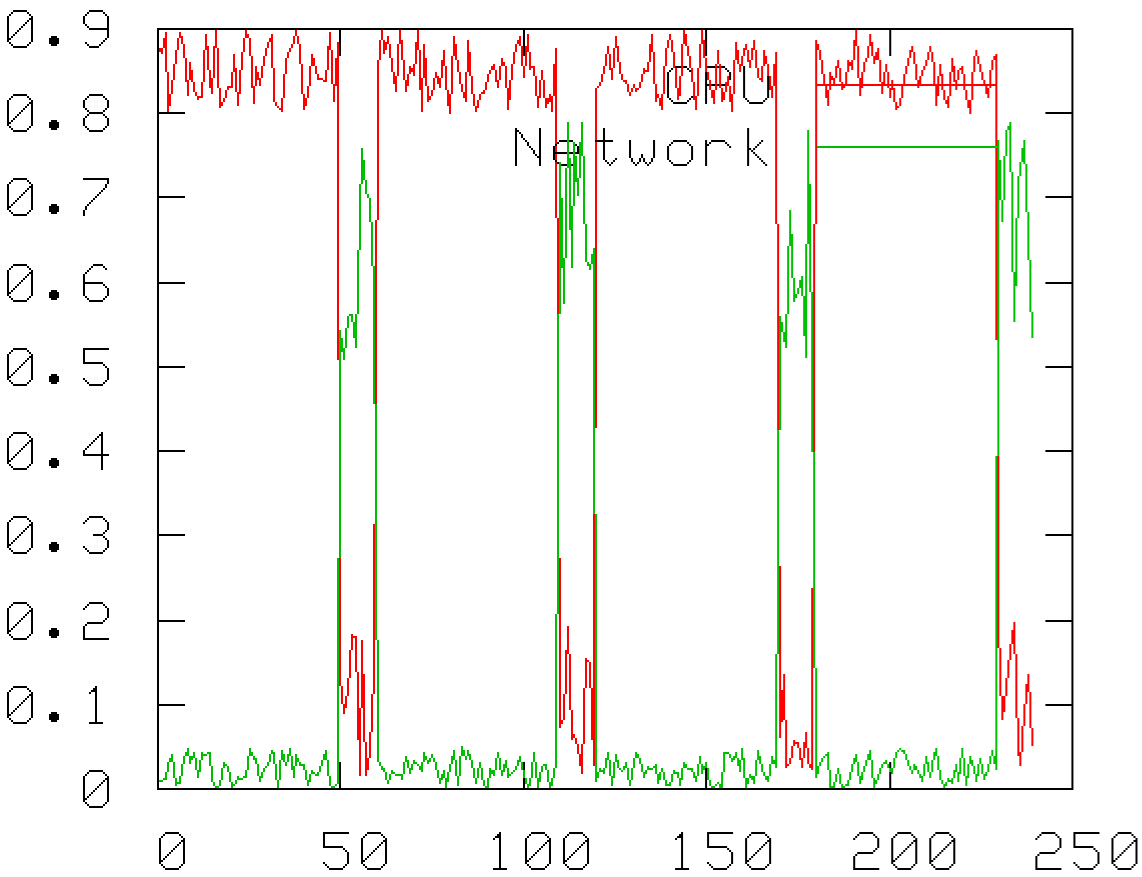
W/N

Weak Scaling





Communication Pattern in Time



Amount sent / bandwidth

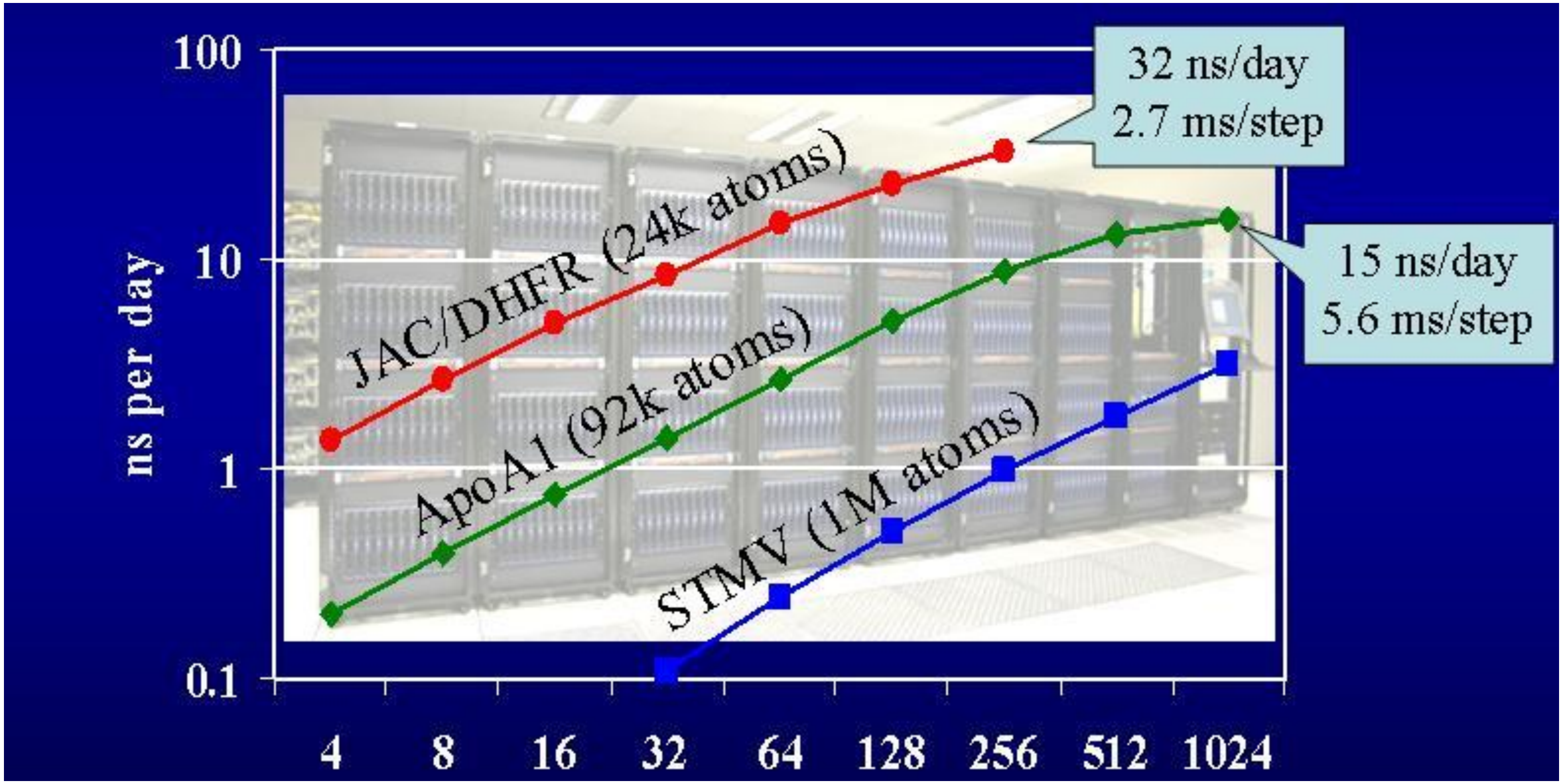
Number of messages × latency

What is the time per iteration?

time = W/N + N/bw + 2 × latency

$$\mathcal{O}(\mathrm{f(N)})$$

It’s LOG!



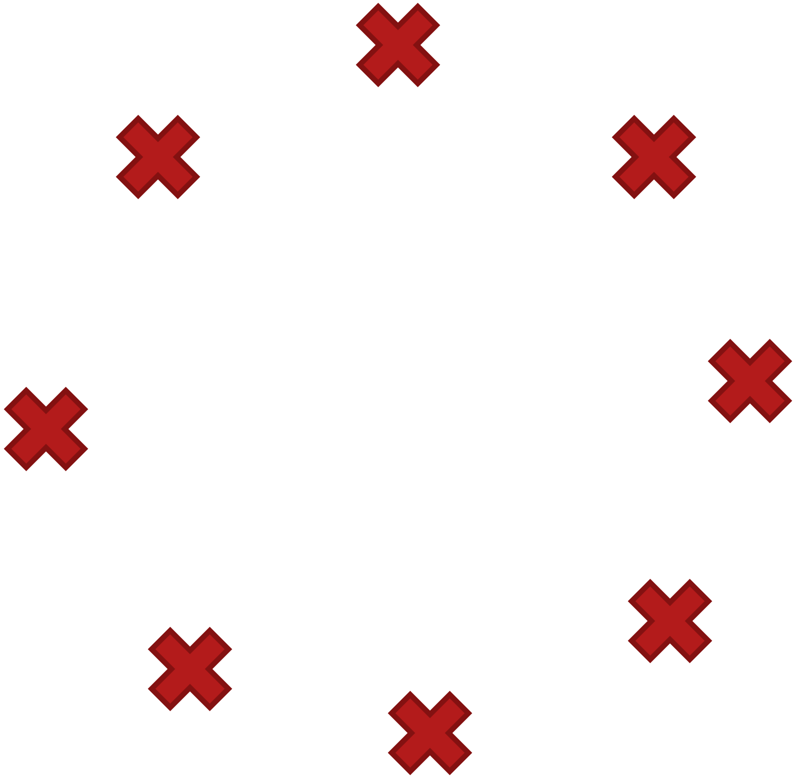
It’s LOG!

Section 4.2 Excel Demo of Fluent
Section 4.3 Measure Strong Scaling

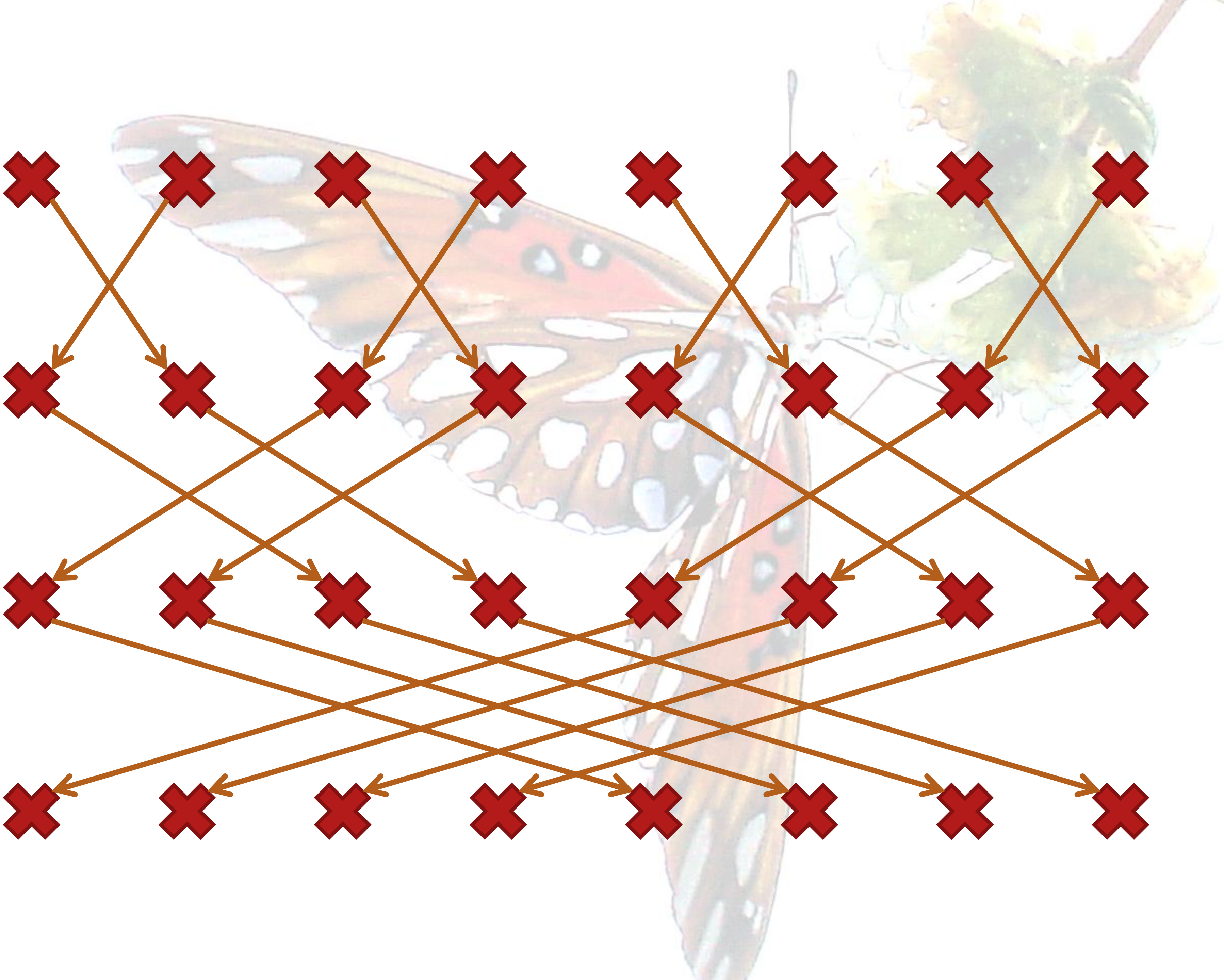
Sun microsystems
Sun microsystems
Sun microsystems
Sun microsystems

2.3 µs

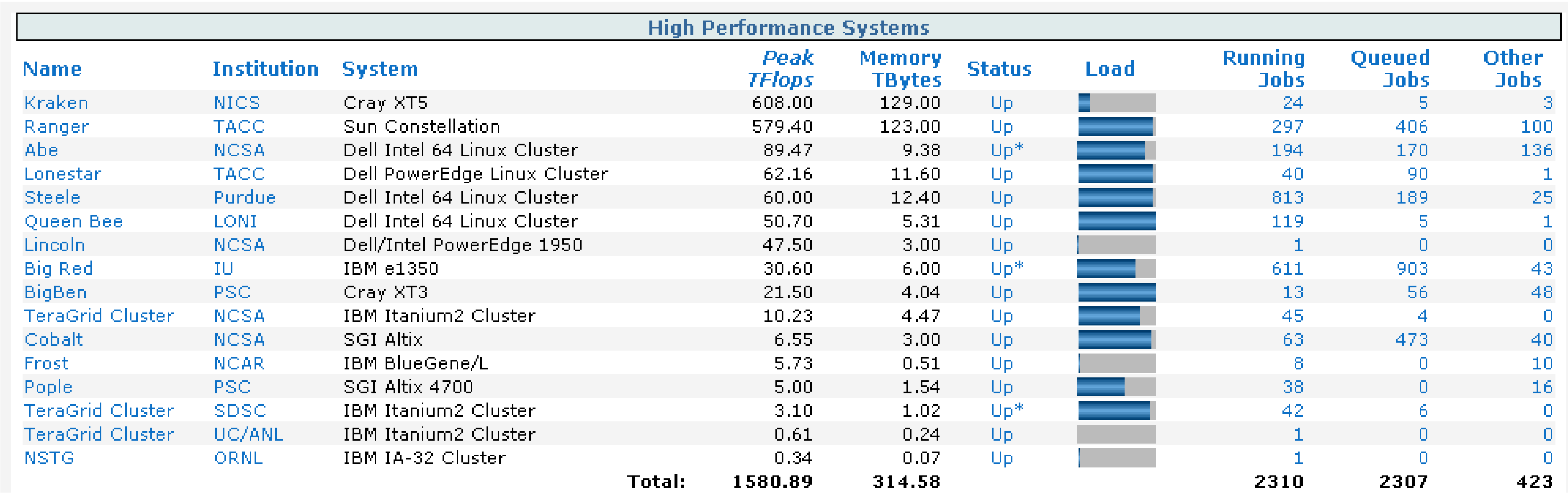
All-to-All Communication



How would you do it? Should you write this yourself?



Right Resources Match Computational Models to Program Models



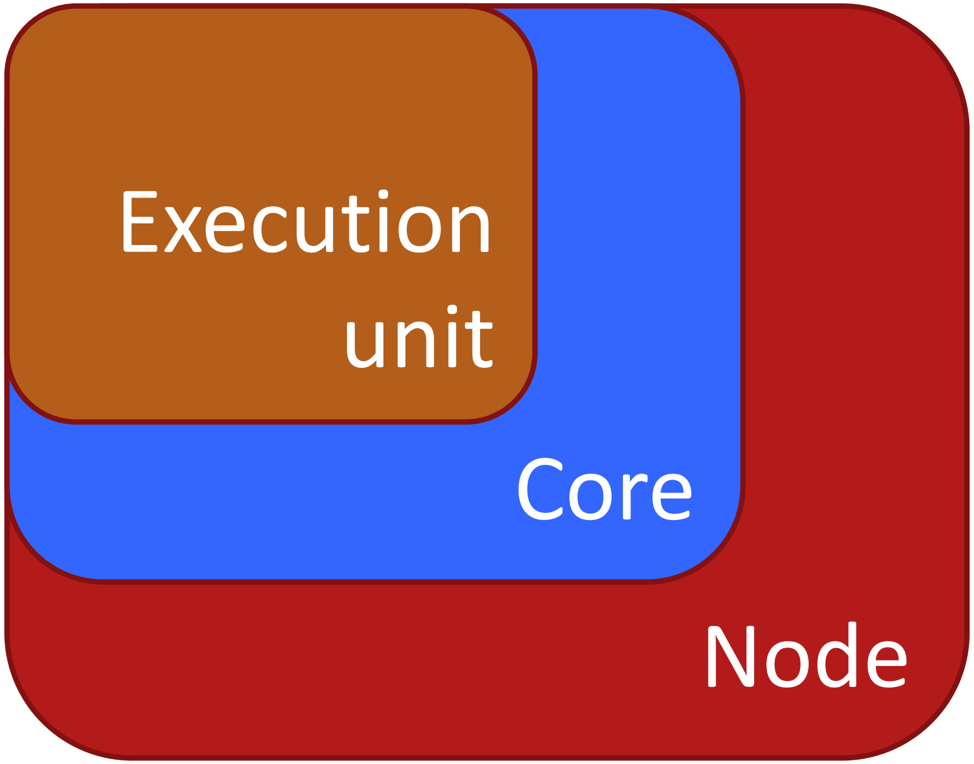
High Performance Systems

Name	Institution	System	*Peak TFlops*	Memory TBytes	Status	Load	Running Jobs	Queued Jobs	Other Jobs
Kraken	NICS	Cray XT5	608.00	129.00	Up		24	5	3
Ranger	TACC	Sun Constellation	579.40	123.00	Up		297	406	100
Abe	NCSA	Dell Intel 64 Linux Cluster	89.47	9.38	Up*		194	170	136
Lonestar	TACC	Dell PowerEdge Linux Cluster	62.16	11.60	Up		40	90	1
Steele	Purdue	Dell Intel 64 Linux Cluster	60.00	12.40	Up		813	189	25
Queen Bee	LONI	Dell Intel 64 Linux Cluster	50.70	5.31	Up		119	5	1
Lincoln	NCSA	Dell/Intel PowerEdge 1950	47.50	3.00	Up		1	0	0
Big Red	IU	IBM e1350	30.60	6.00	Up*		611	903	43
BigBen	PSC	Cray XT3	21.50	4.04	Up		13	56	48
TeraGrid Cluster	NCSA	IBM Itanium2 Cluster	10.23	4.47	Up		45	4	0
Cobalt	NCSA	SGI Altix	6.55	3.00	Up		63	473	40
Frost	NCAR	IBM BlueGene/L	5.73	0.51	Up		8	0	10
Pople	PSC	SGI Altix 4700	5.00	1.54	Up		38	0	16
TeraGrid Cluster	SDSC	IBM Itanium2 Cluster	3.10	1.02	Up*		42	6	0
TeraGrid Cluster	UC/ANL	IBM Itanium2 Cluster	0.61	0.24	Up		1	0	0
NSTG	ORNL	IBM IA-32 Cluster	0.34	0.07	Up		1	0	0
		Total:	**1580.89**	**314.58**			**2310**	**2307**	**423**

- All more complex than what we have described.
- Include RAM, flash, disk, tape, WAN.

Profiling and Presents

Speeds and Feeds at Scales



http://www.flickr.com/photos/35188692@N00/83775147/

http://www.flickr.com/photos/avdleeuw/48388892/