MPI Network Performance

- Scalability
- Latency
- Bandwidth

Example

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} e \\ f \end{bmatrix} = \begin{bmatrix} a \cdot e + b \cdot f \\ c \cdot e + d \cdot f \end{bmatrix}$$

CPU 1

CPU 2

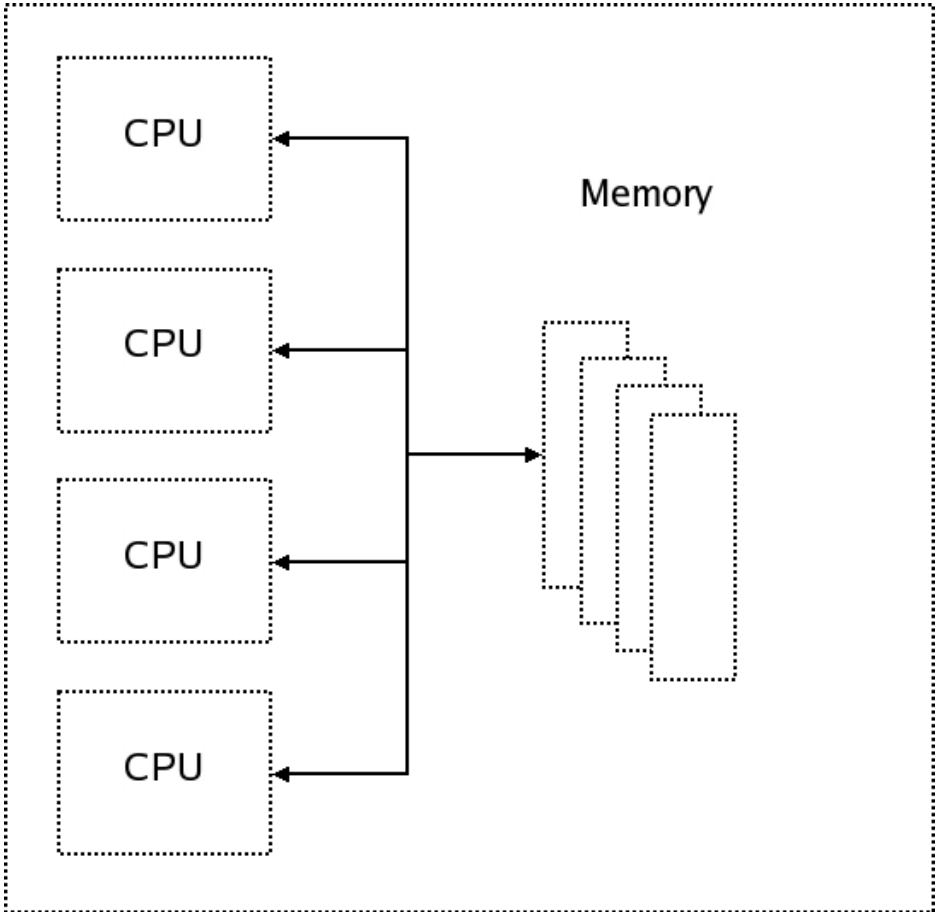
$$M\vec{b} = \vec{C}$$

Many numerical methods use matrix calculation and can be parallelized.

- BLAS -> ATLAS -> Pthread

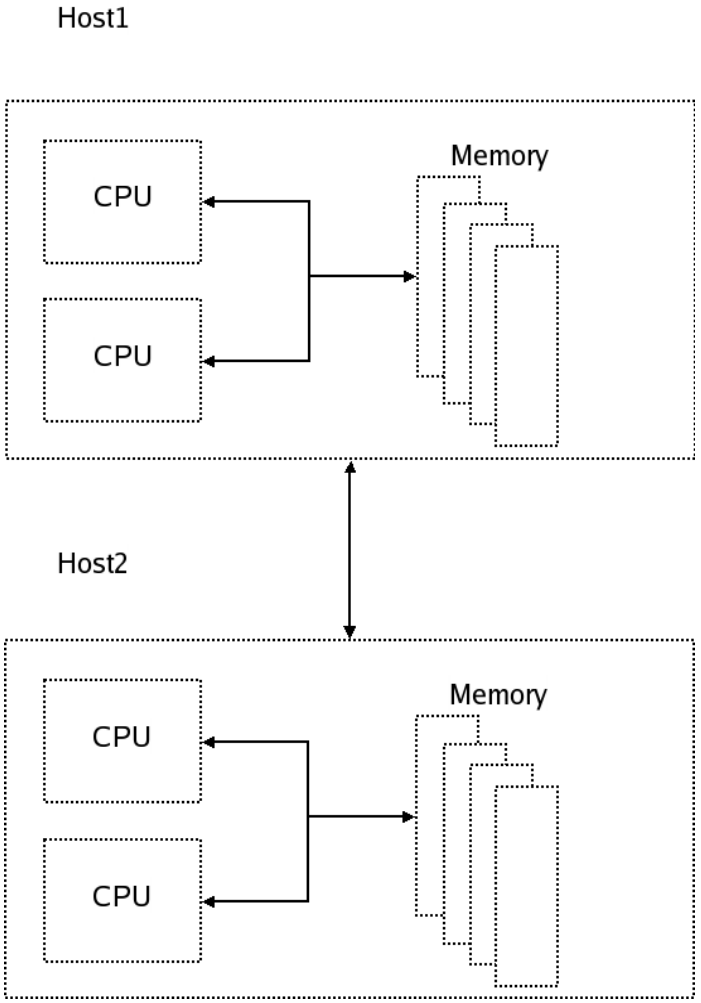
Parallel Approaches

- Posix Threads
  - Well understood
  - Shared Memory
  - Simple Mutexs
  - Not Cheap



Parallel Approaches

- MPI (Message Passing Interface)
  - Shared or distributed Memory
  - Well supported
  - Portable
  - Explicit Data Passing

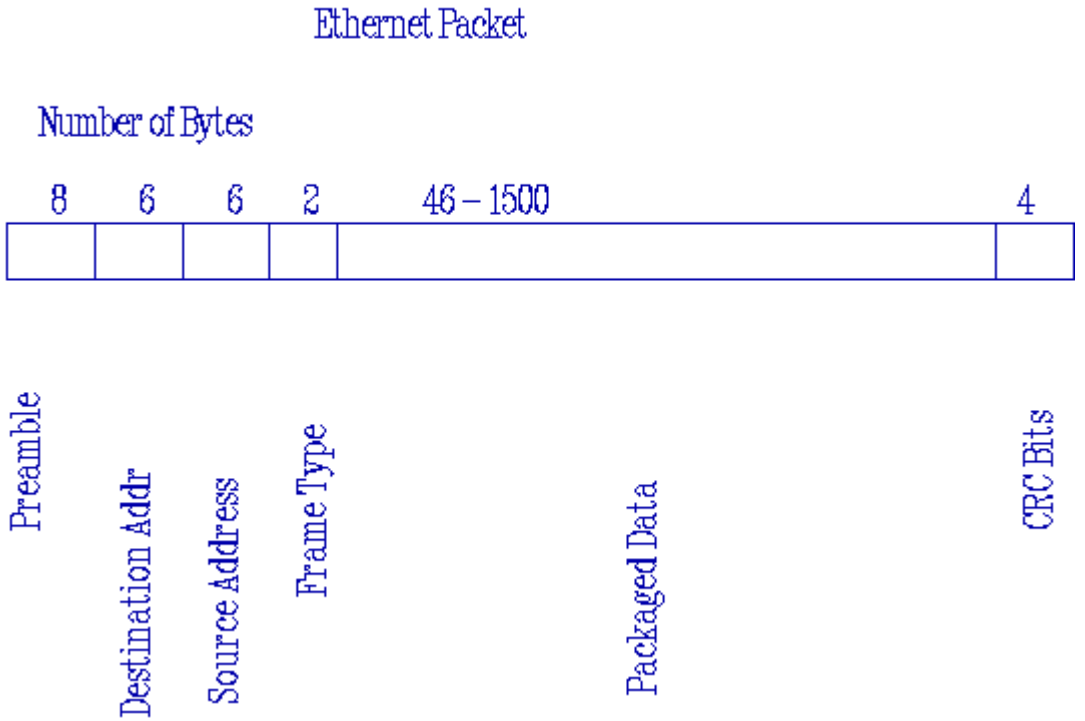


The Networks

- Myrinet 2000
  - 2Gb/s
  - Uses GM driver
- Ethernet
  - 1Gb/s
  - Jumbo Frames

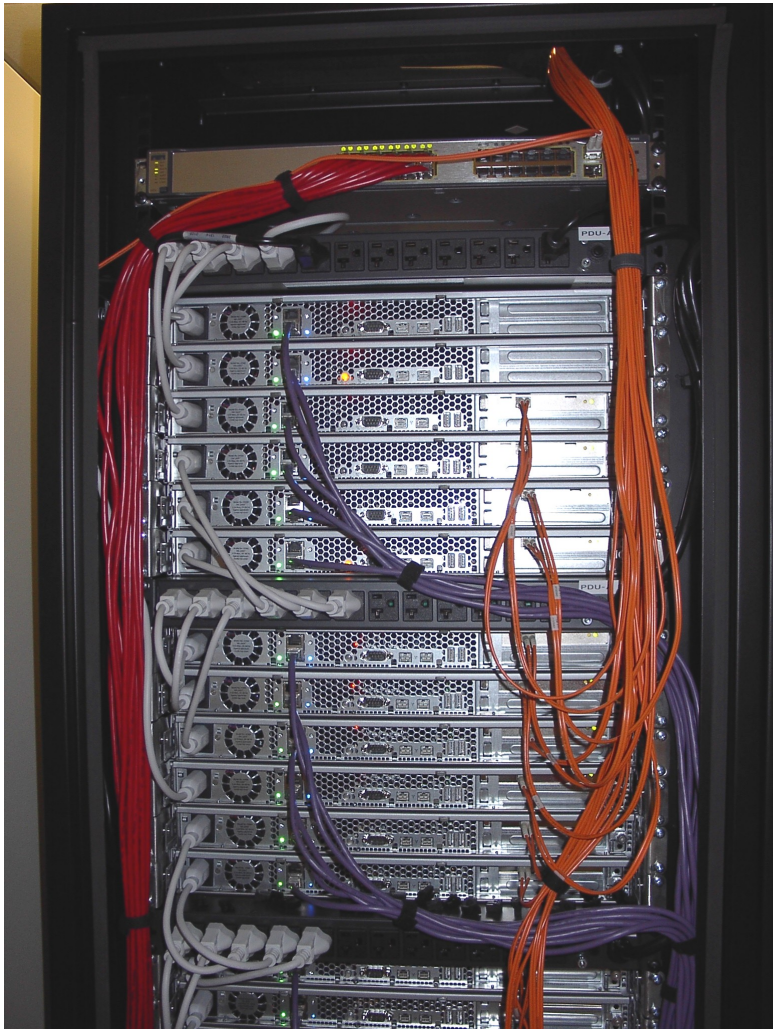
Ethernet

- Cheap
- Reliable
- Jumbo Frames
- Slow
- TCP/IP



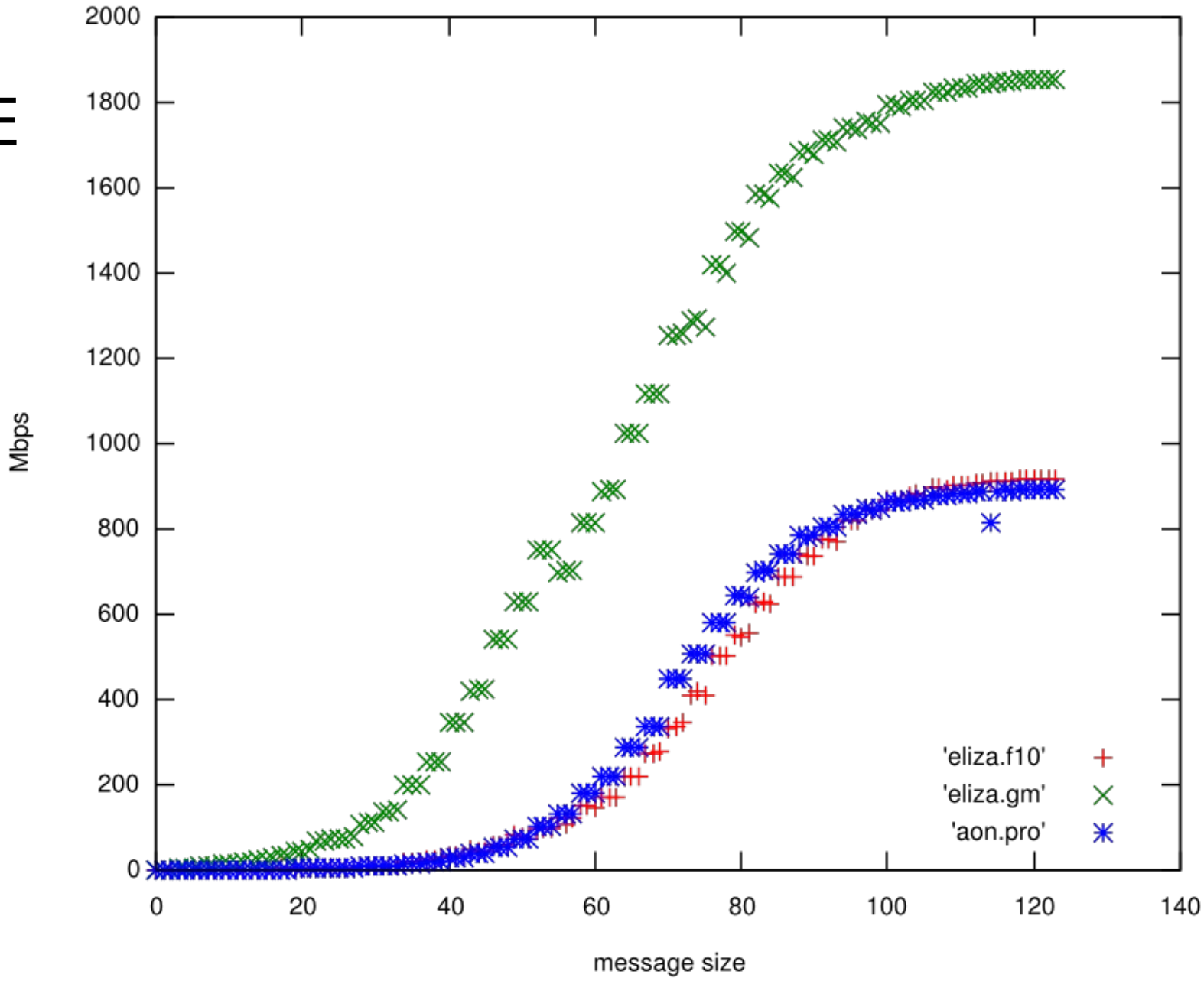
Myrinet

- Fast (For Now)
- No TCP/IP
- Well Supported

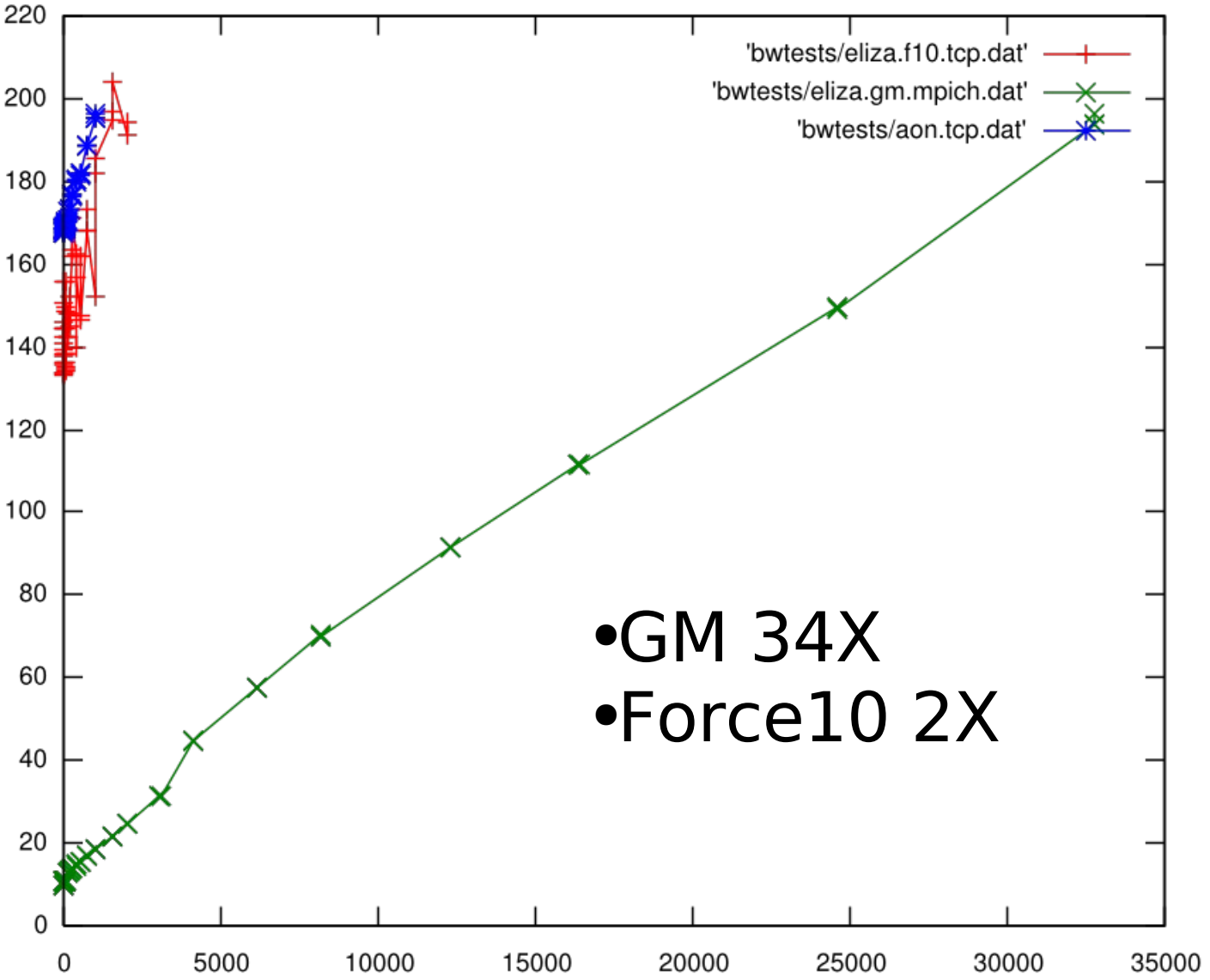


Bandwidth

- NetPIPE
- TCP/IP



Latency

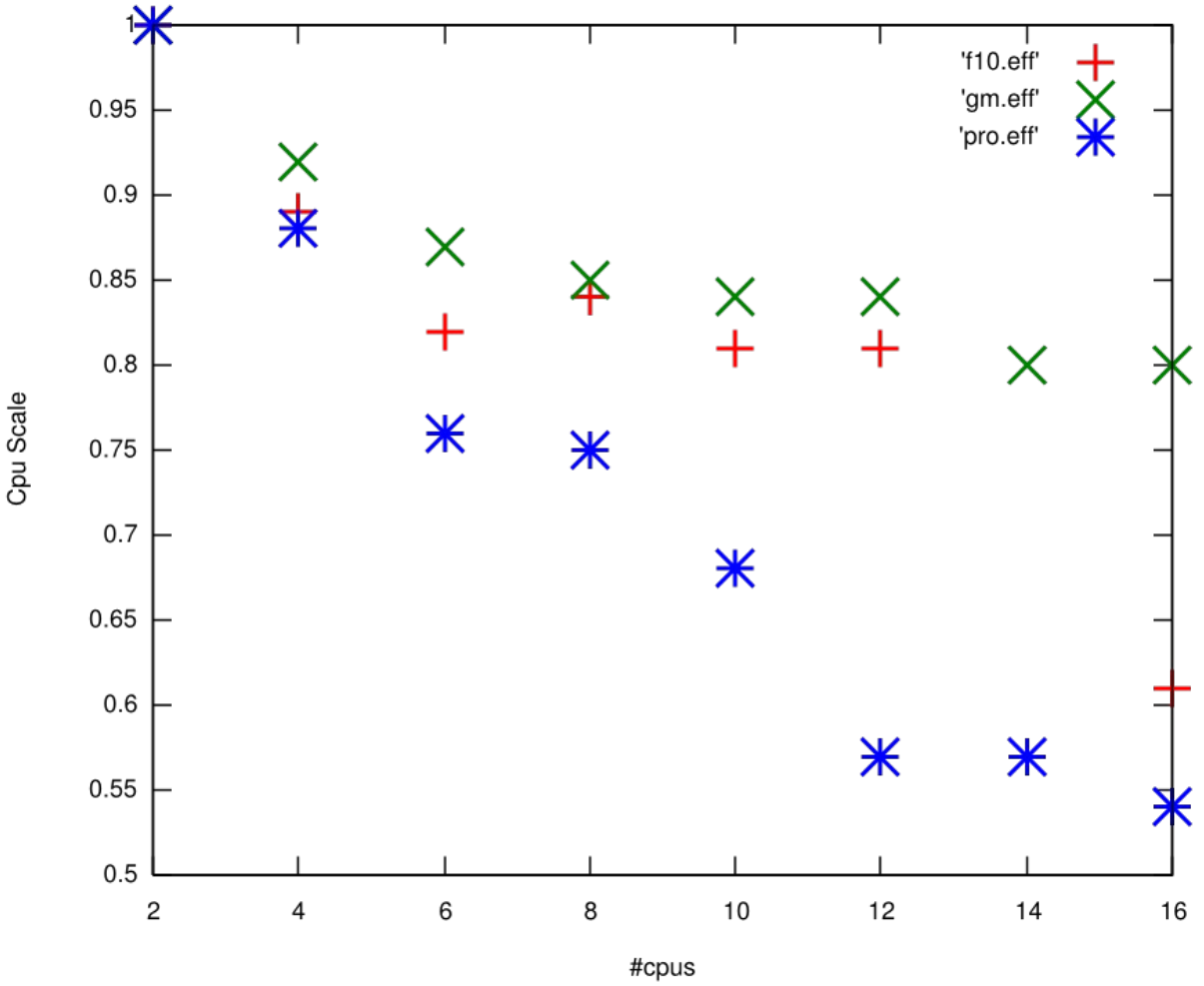


- GM 34X
- Force10 2X

Cpu Scaling

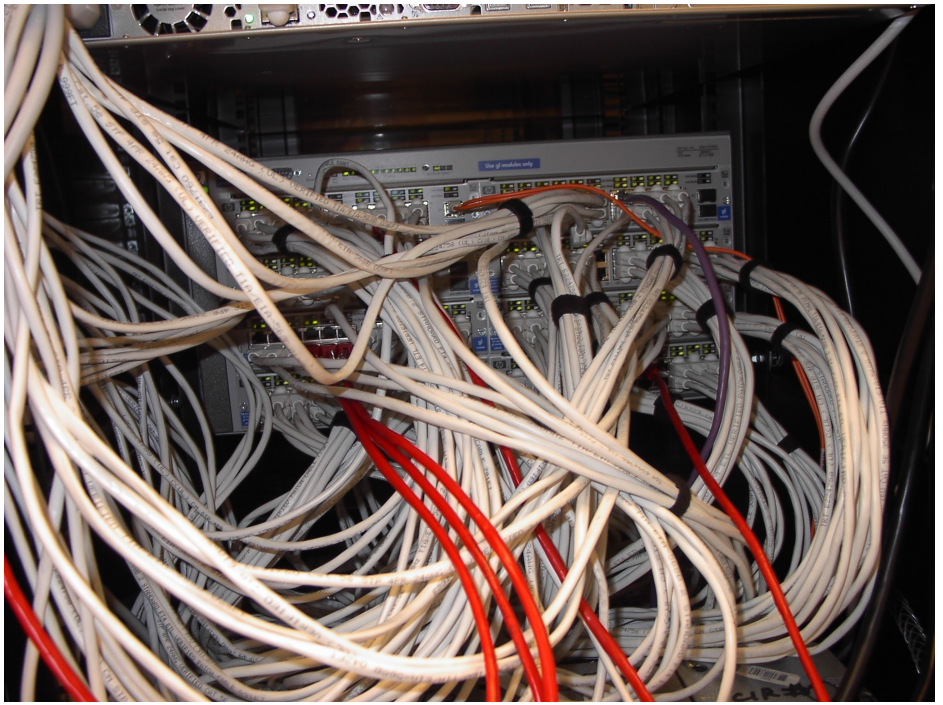
$$f_c N_{fpu} \xi = P_{fpu}$$

$$\epsilon = \frac{P_w}{N_{node}\left(P_{mfpu} N_{cpu}\right)}$$



Recommendations

- Embarrassingly Parallel
  - MCNP5
  - Seti@home
- Tightly Coupled
  - Boundary Condition
  - HPL



Checklist

- Problem Run Time
- Problem Nature
- Cost
- Shared System
- Do you *NEED* Shared Memory?

Who are we?

- 584 Nodes (1,168 CPU's)
- 1,244 GB RAM
- 11 TB Shared Disk
- 30 TB Scratch
- 0.58 Tb/s Network
- 4 Clusters 3 Platforms 2 OS's


The Center for Advanced Computing
The University of Michigan