UCIrvine | The Henry Samueli School of Engineering

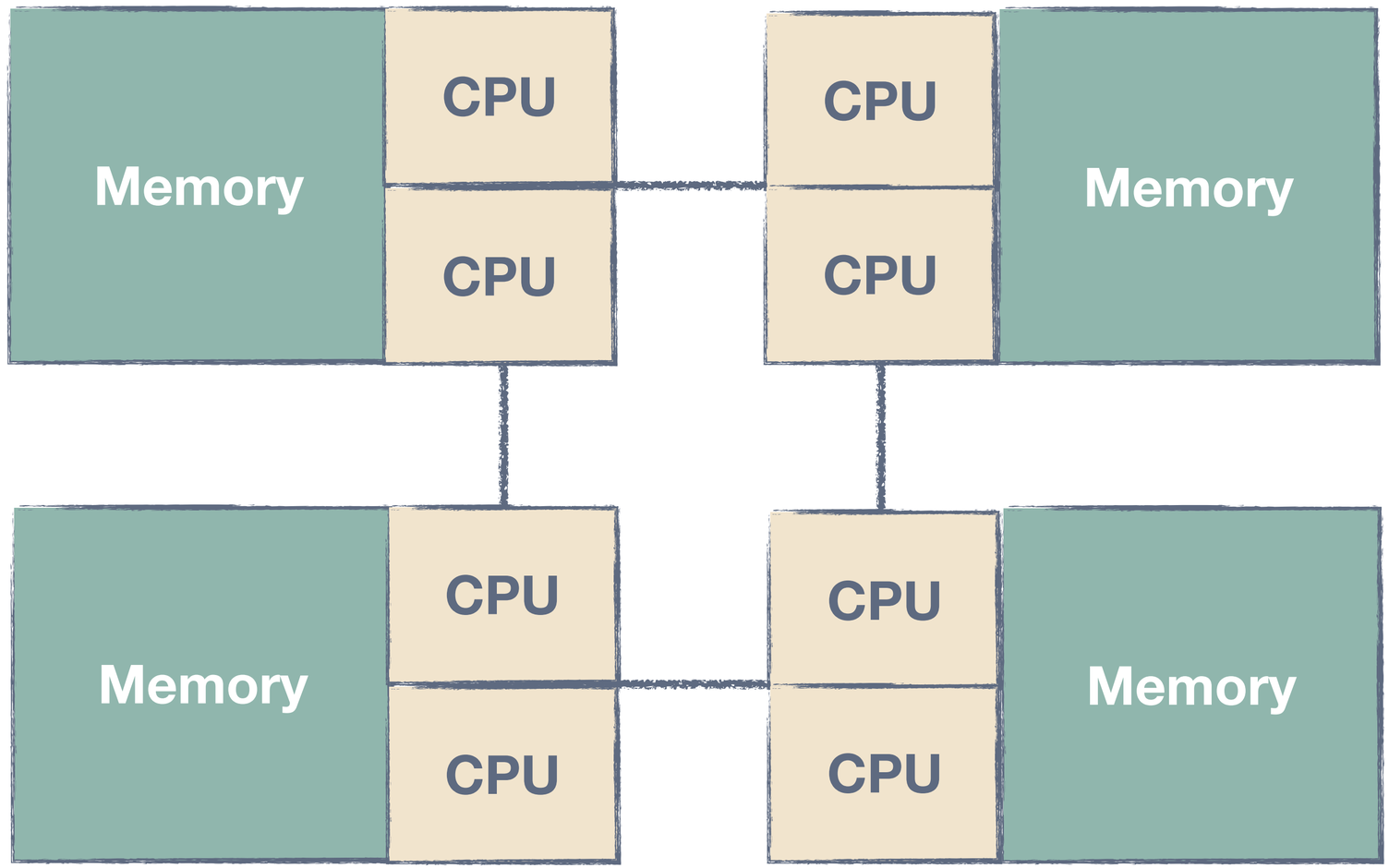
Distributed memory model and algorithms

EECS 221: Intro to High-Performance Computing

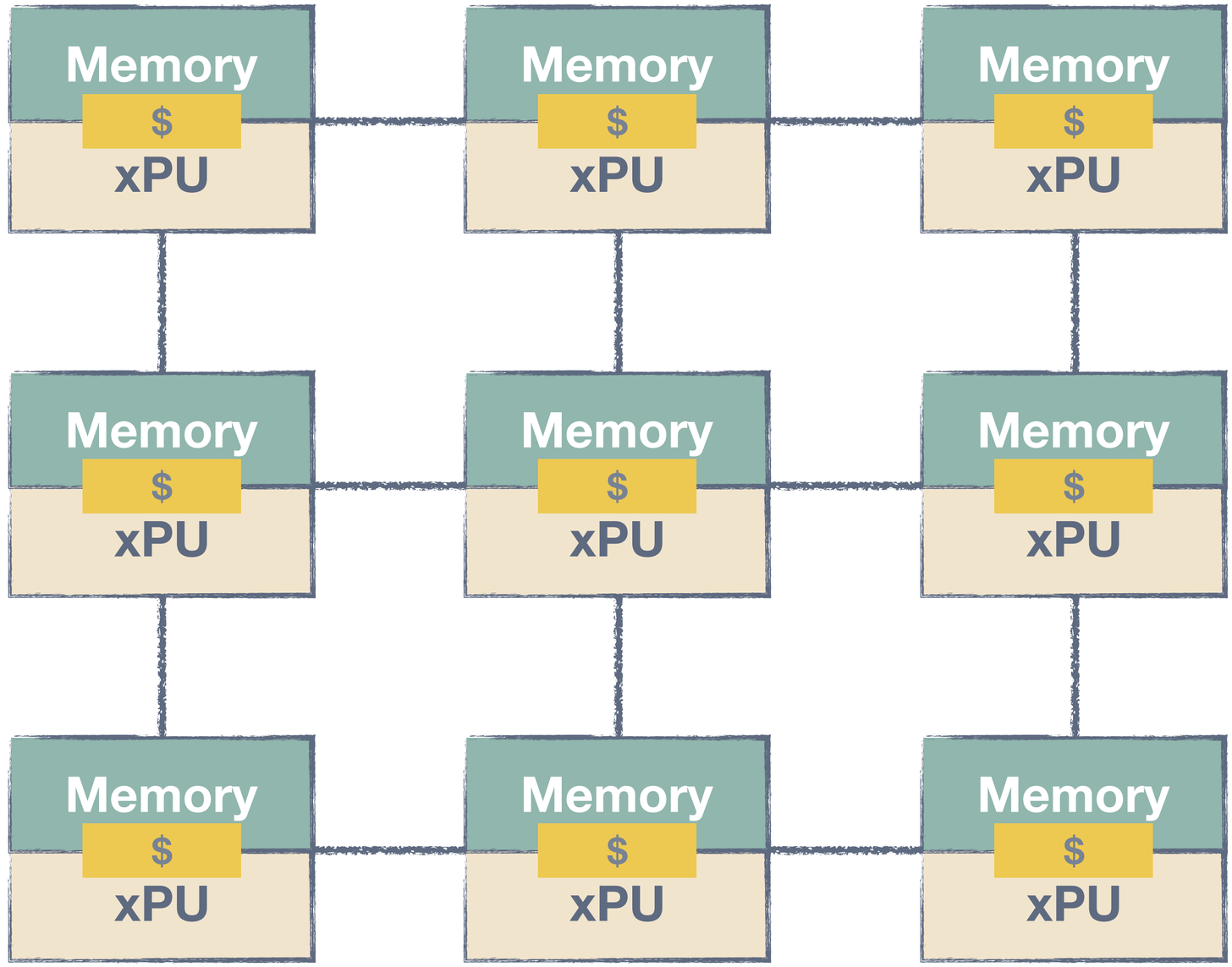
Aparna Chandramowlishwaran

April 20, 2017

Recall: Non-uniform memory access



Example: Four dual-core CPU's with logically shared but physically distributed memory



Distributed Memory System

Each processor has its own private memory and cache but cannot directly access another processor's memory. Communicate over a network using message passing.

Recall: Amdahl's Law

- Computation time on a sequential machine

$$T_1 = t_s + t_p$$

- Computation time on a parallel machine assuming ideal parallelization

$$T_p = t_s + \frac{t_p}{p}$$

- **t_s** = Computation time for the sequential part
- **t_p** = Computation time for the parallelizable part
- **p** = Number of processors

Recall: Amdahl's Law

- Computation time on a sequential machine

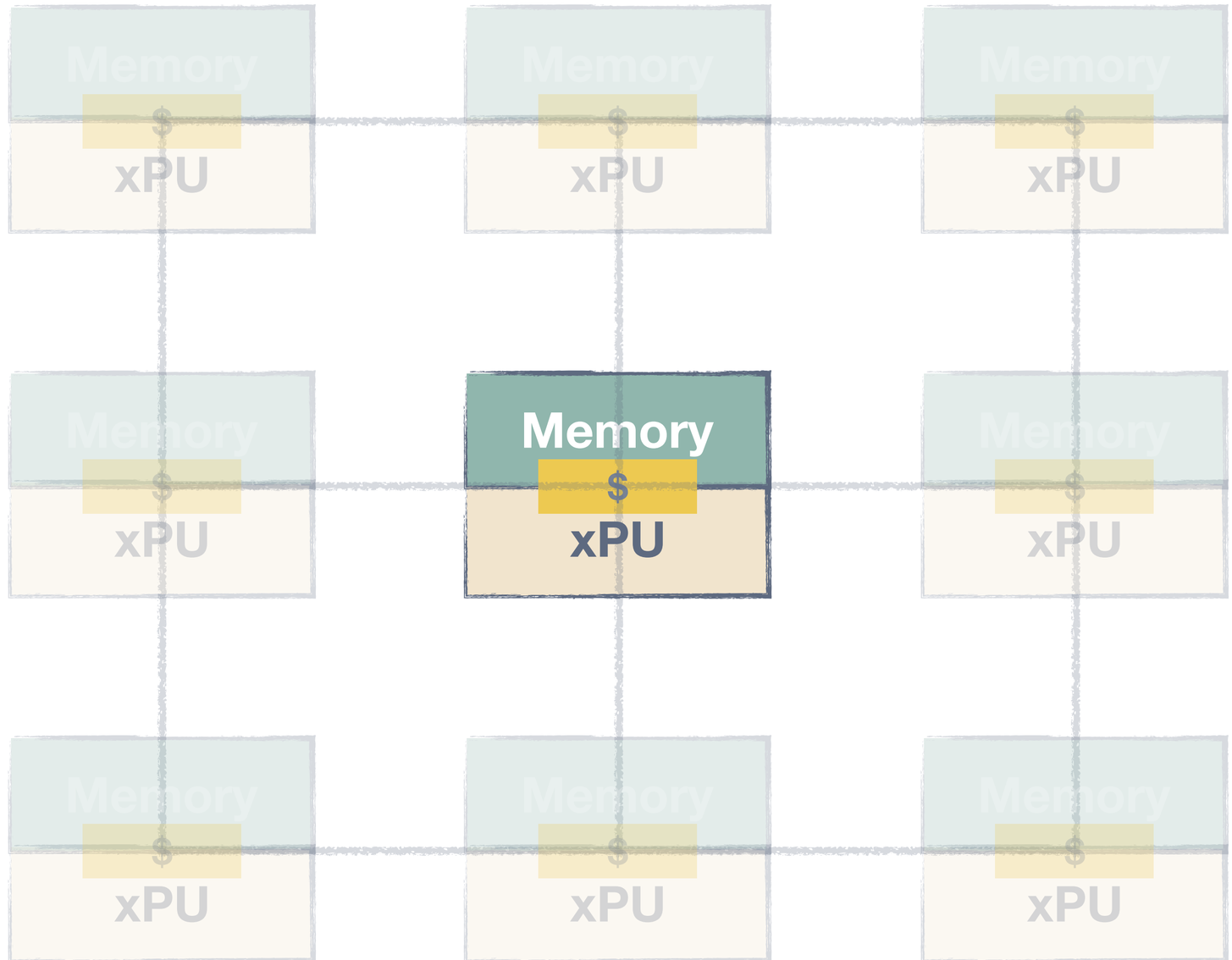
$$T_1 = t_s + t_p$$

- Computation time on a parallel machine assuming ideal parallelization

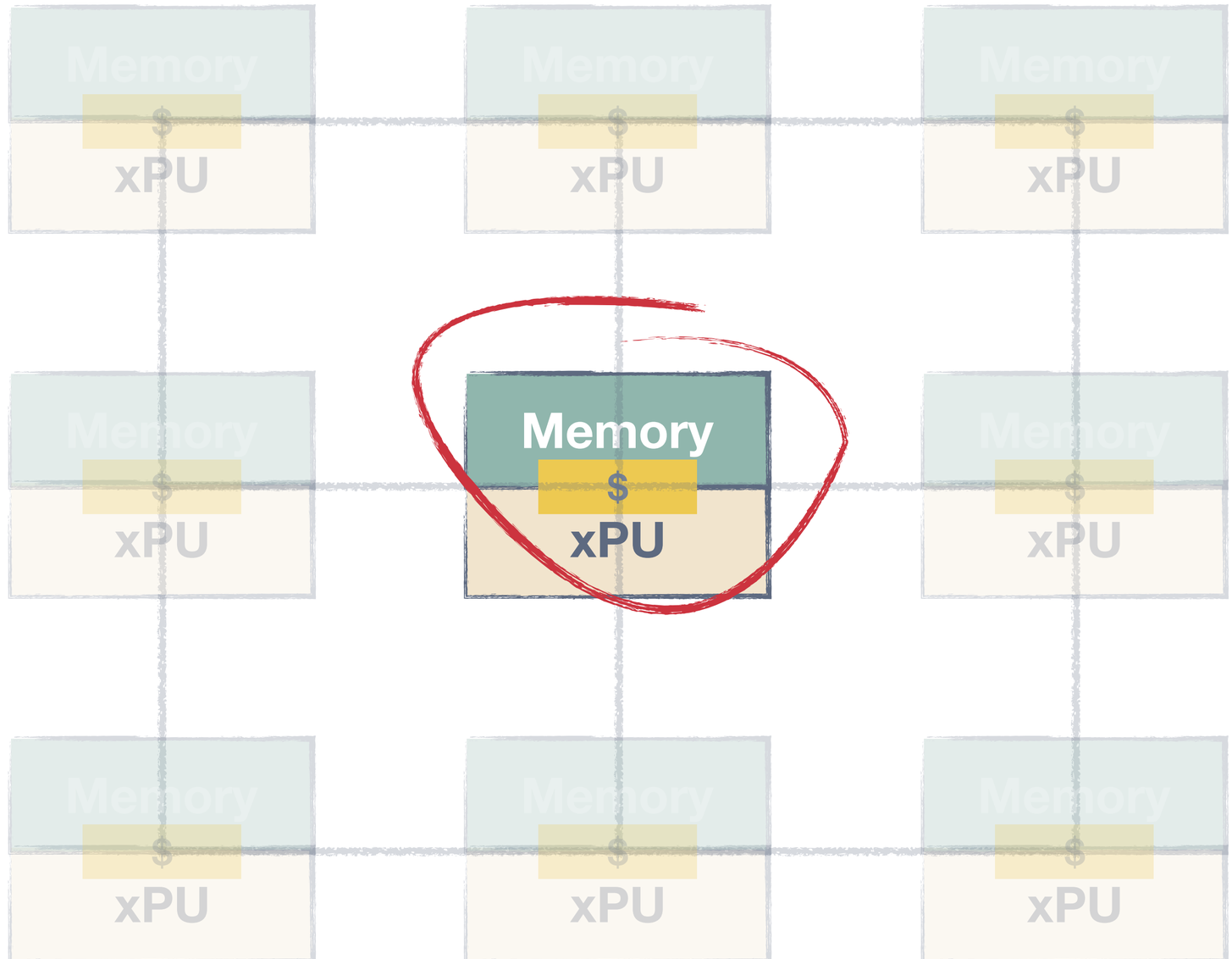
$$T_p = t_s + \frac{t_p}{p}$$

- In practice

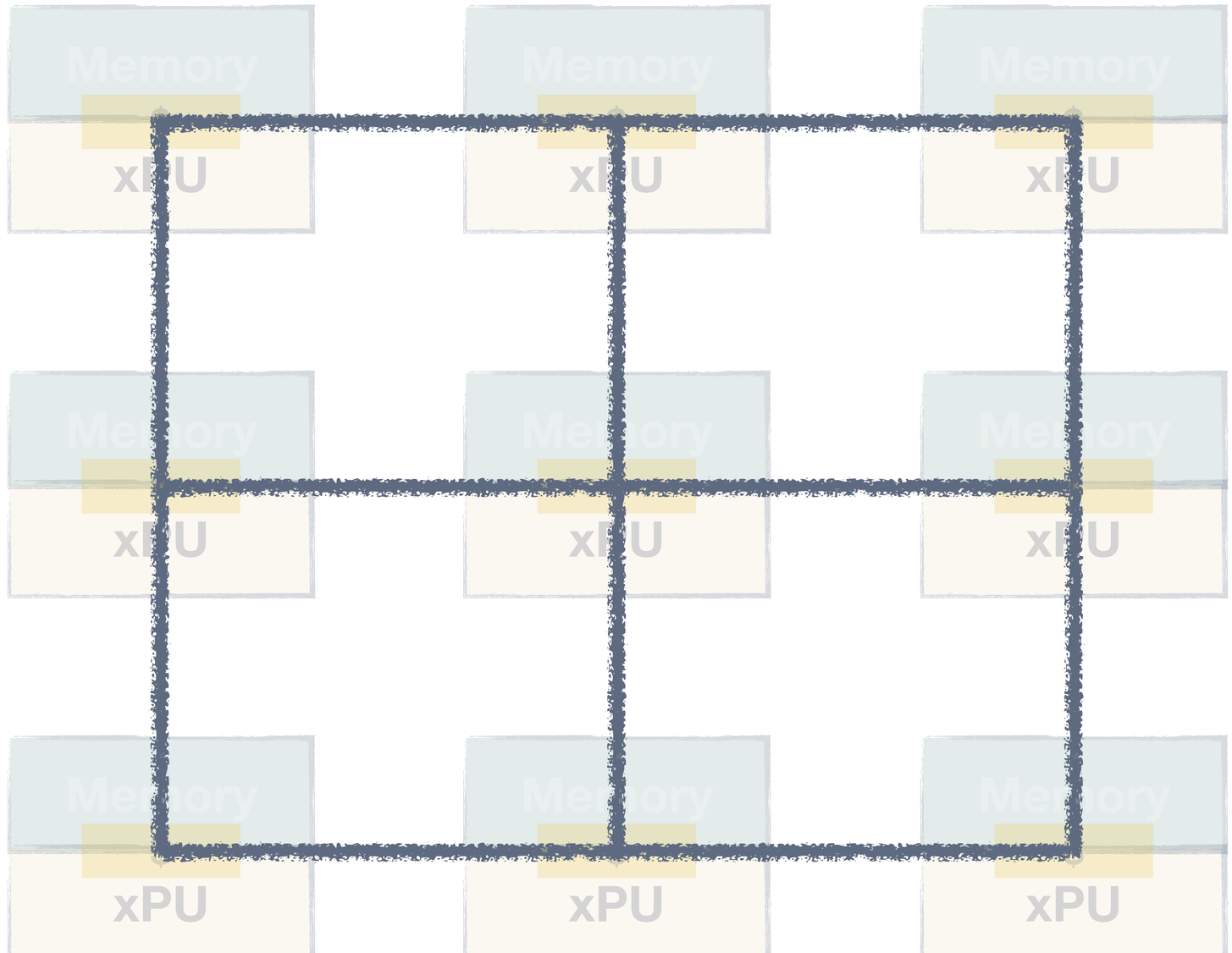
$$T_p = t_s + \frac{t_p}{p} + T_{comm} \quad \text{(or)} \quad T_p = T_{comp} + T_{comm}$$



Two costs: $T_{\text{network}} + T_{\text{memory}}$



Two costs: $T_{\text{network}} + T_{\text{memory}}$



Two costs: $T_{network} + T_{memory}$

Cost model for distributed memory algorithms

“α-β” (latency-bandwidth) cost model

- Time to send a message of size **n**

$$T(n) = \alpha + \frac{n}{\beta}$$

- Usually, α >> 1/β >> cost(flop)
  - One long message is cheaper than many short ones
  - Can do thousands of flops for each message
- Lesson: Need large computation-to-communication ratio to be efficient
- LogP — more detailed models exist (**L**atency/**o**verhead/**g**ap/**P**rocessors)

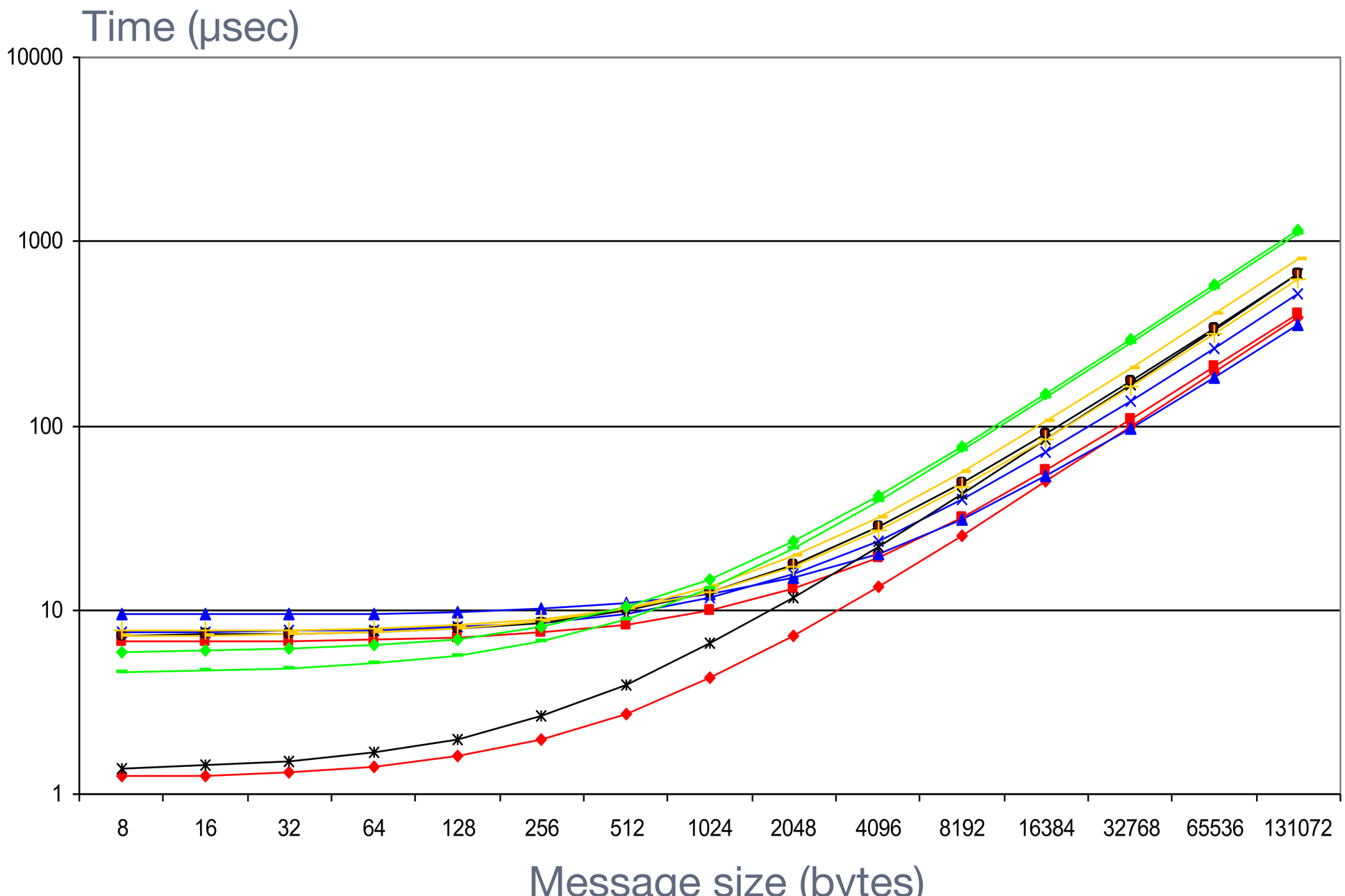
Latency and (inverse)bandwidth on real machines

machine	α	β
T3E/Shm	1.2	0.003
T3E/MPI	6.7	0.003
IBM/LAPI	9.4	0.003
IBM/MPI	7.6	0.004
Quadrics/Get	3.267	0.00498
Quadrics/Shm	1.3	0.005
Quadrics/MPI	7.3	0.005
Myrinet/GM	7.7	0.005
Myrinet/MPI	7.2	0.006
Dolphin/MPI	7.767	0.00529
Giganet/VIPL	3.0	0.010
GigE/VIPL	4.6	0.008
GigE/MPI	5.854	0.00872

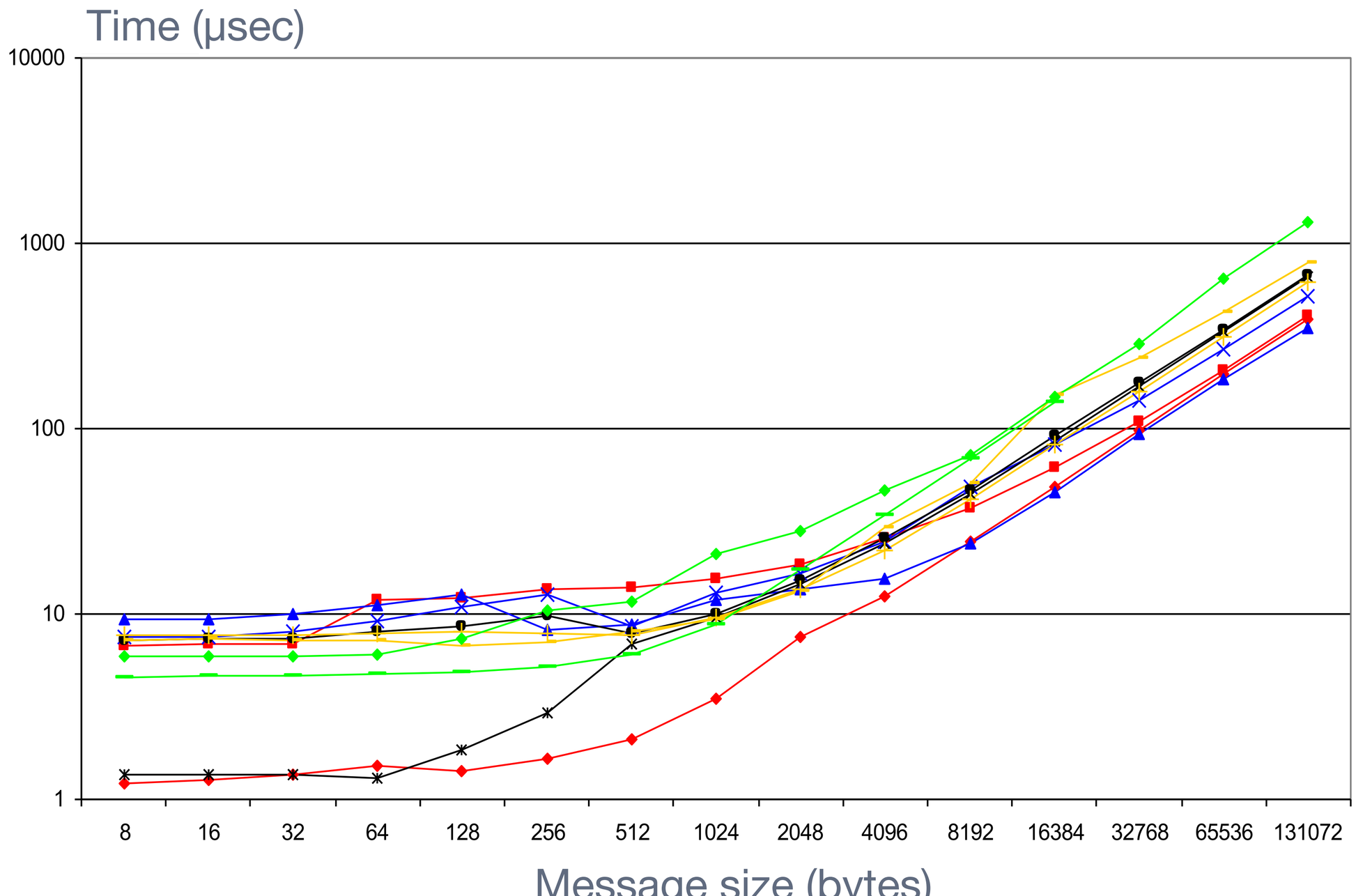
Data obtained empirically.

α in μs
β^{-1} in μs/byte

Time to send a message (α-β model)



Time to send a message (measured)



Example: Distributed matrix multiply (whiteboard)