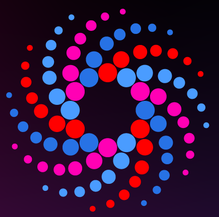
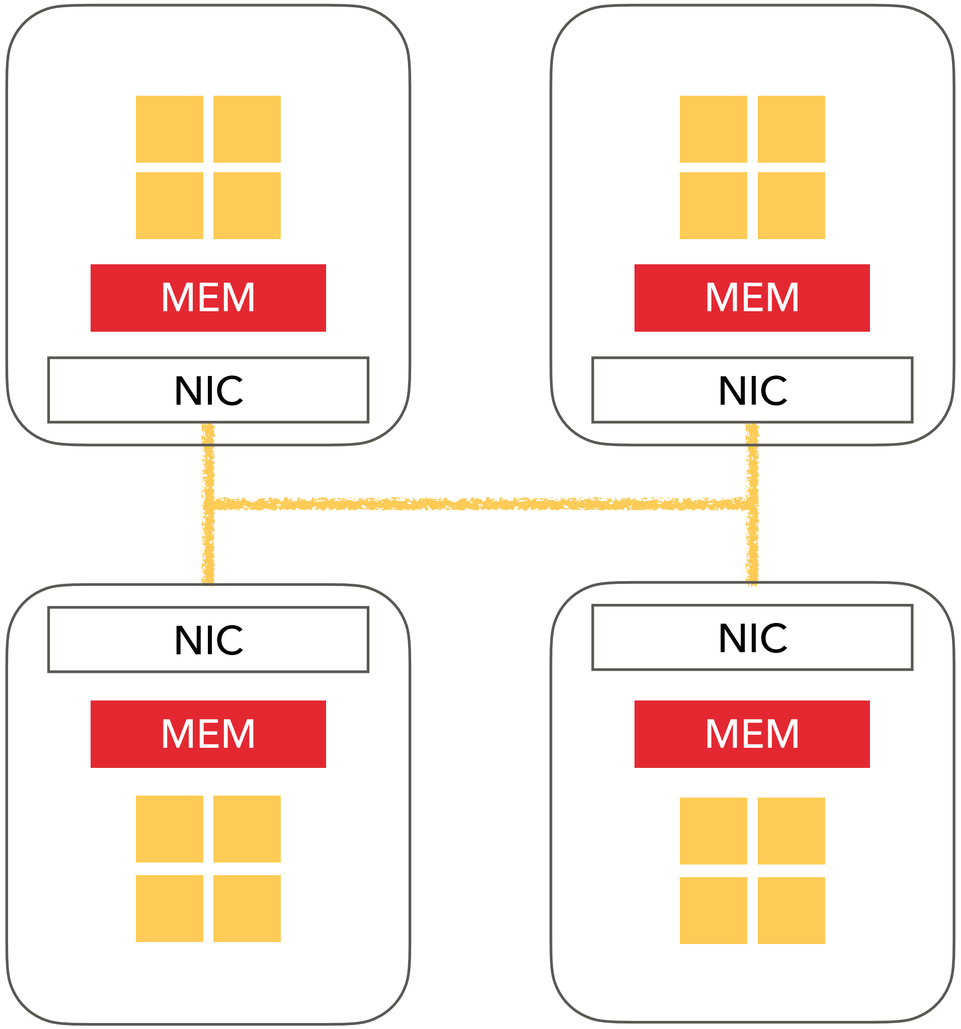


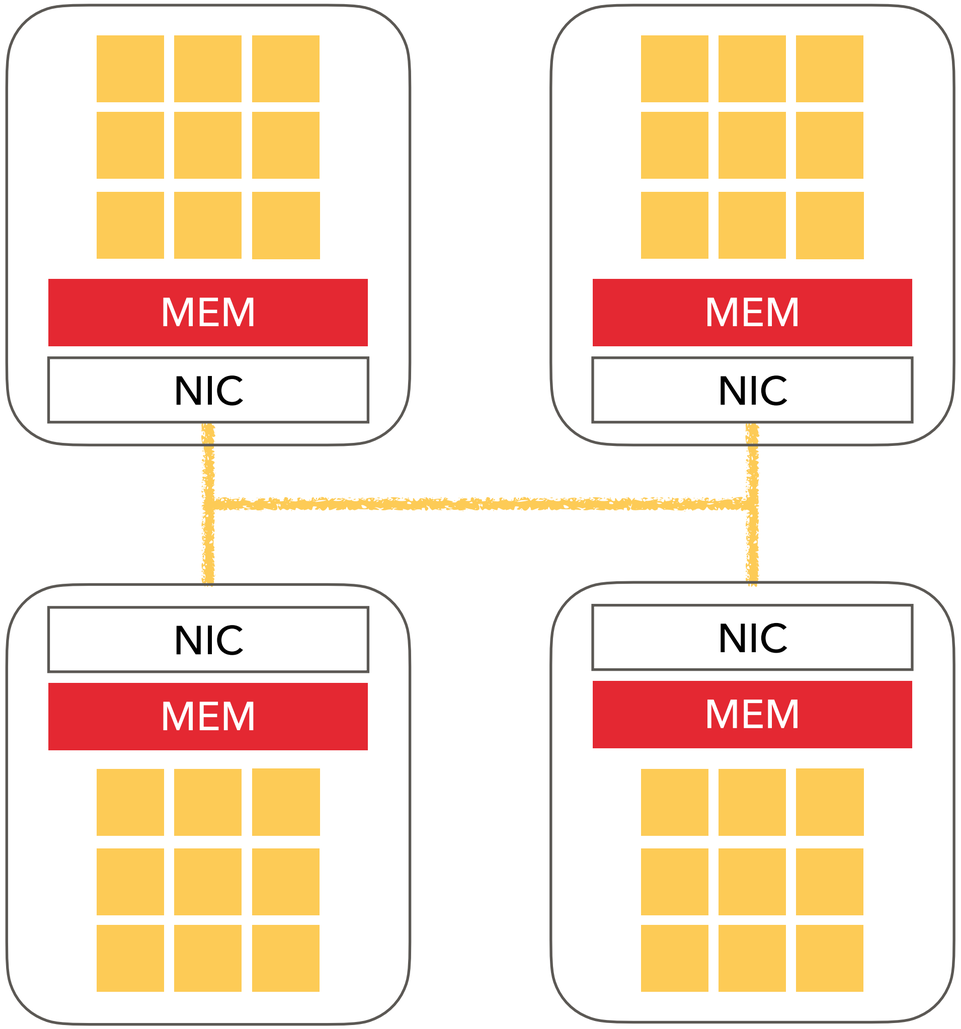
SC22
Dallas, TX | hpc accelerates.

Network Communication in Heterogeneous Computing Environments

Rohit Zambre (Senior Software Architect, Nvidia)


MEM
NIC
MEM
NIC
NIC
MEM
NIC
MEM

- *Trend 1*: Disproportionate increase in number of cores per processor



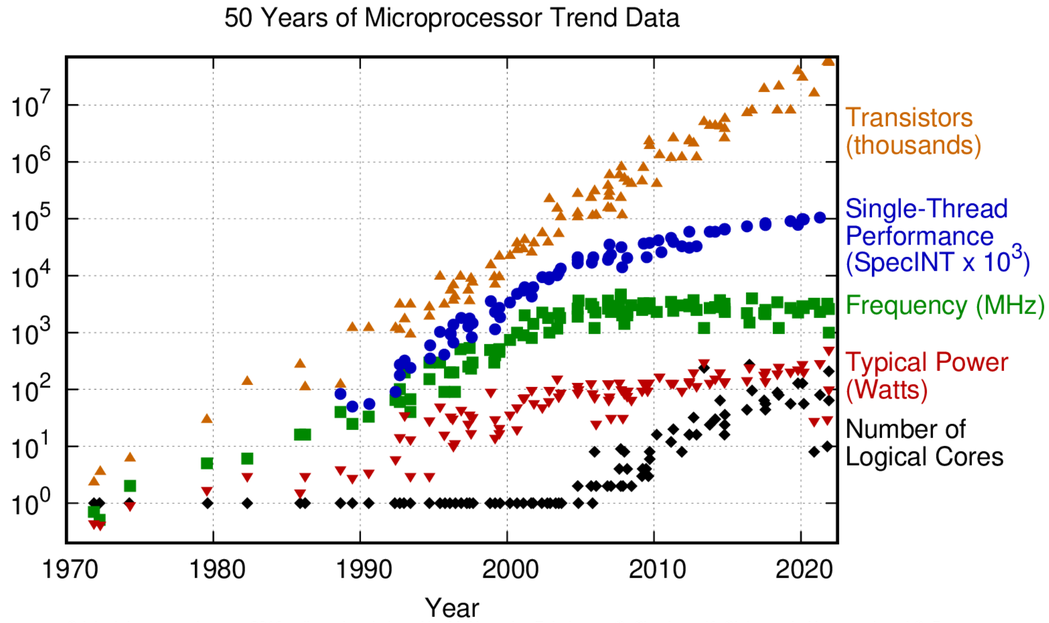
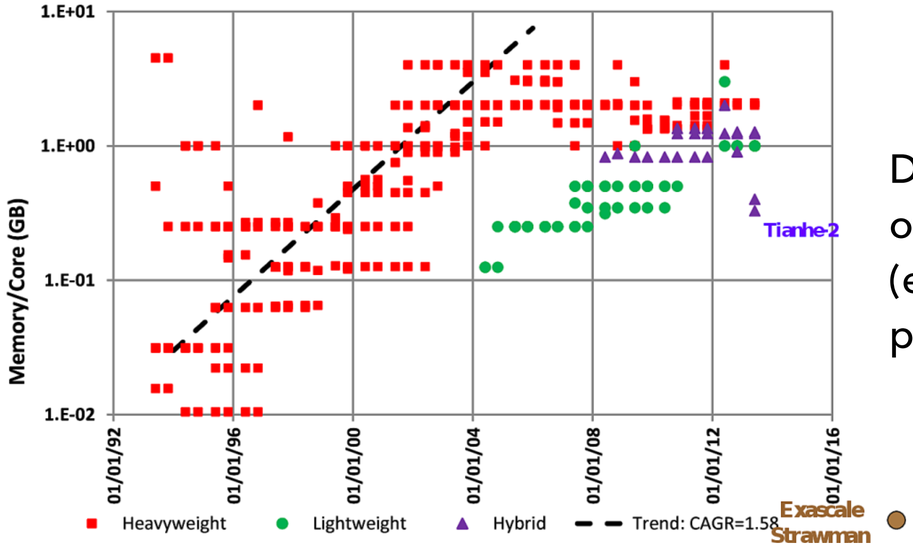


Image: https://github.com/karlrupp/microprocessor-trend-data

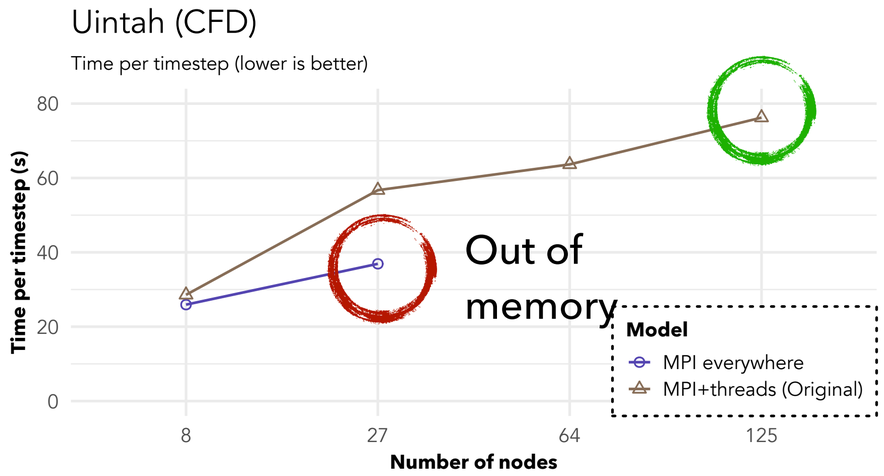
Increasing number of cores

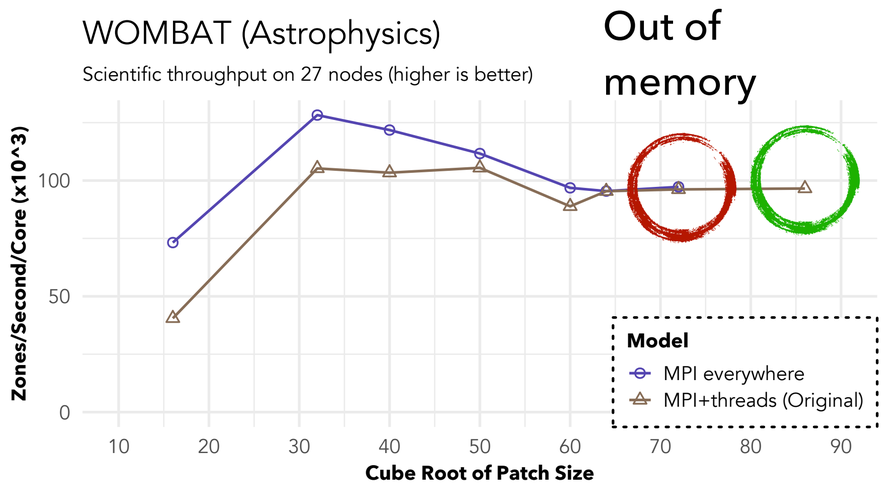


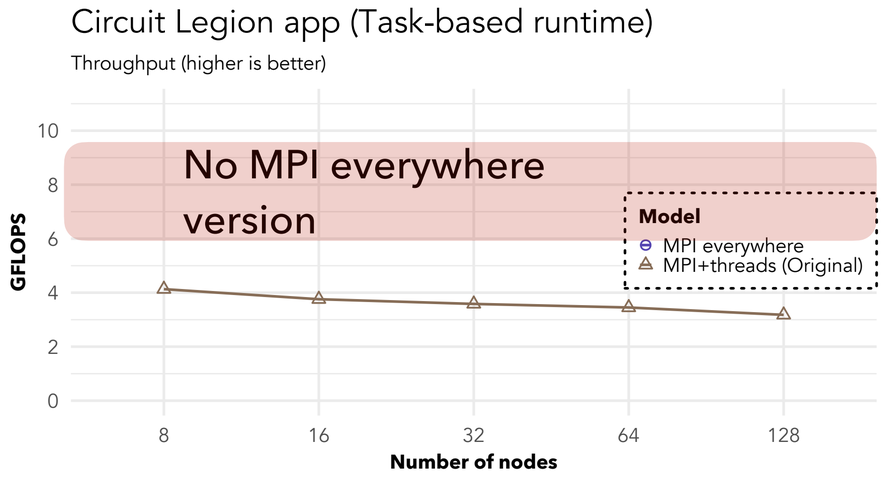
Decreasing share of resources (e.g., memory) per core

MPI+Threads for scalability

MPI everywhere has scalability issues



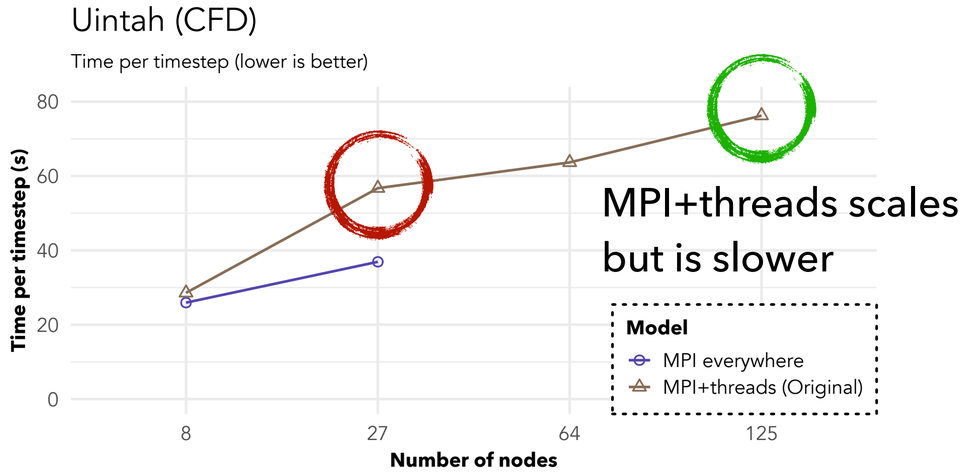


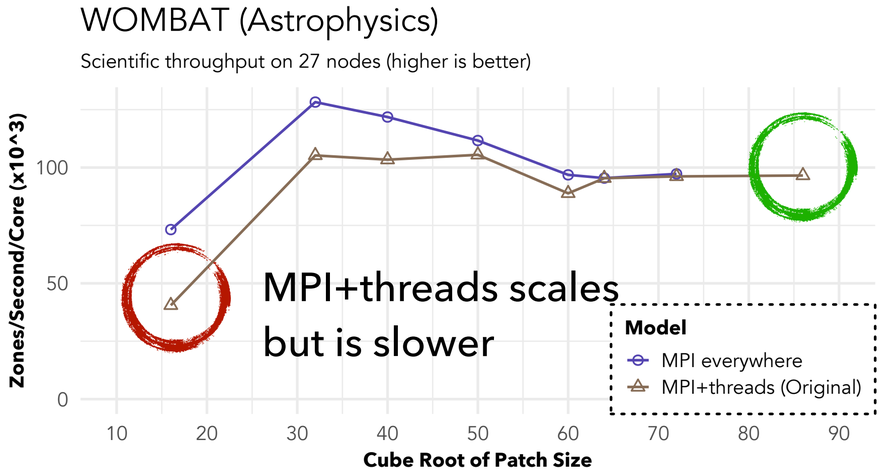


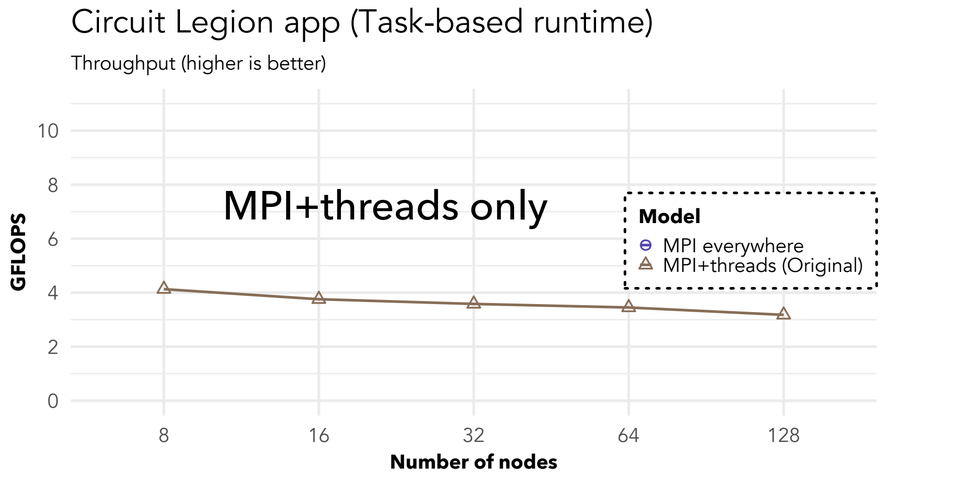
- MPI+threads able to scale on modern architectures

MPI+Threads for scalability

But MPI+Threads has performance hurdles







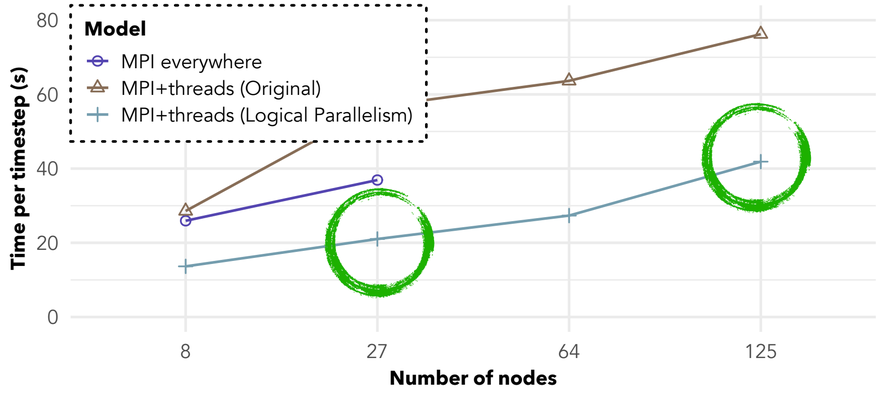
- MPI+threads able to scale on modern architectures
- MPI+threads poses many challenges

MPI+Threads for scalability and performance

Key: logically parallel communication

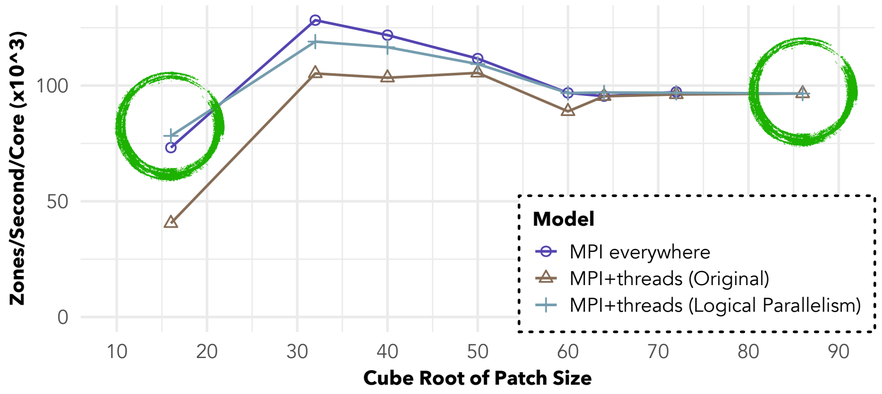
Uintah (CFD)

Time per timestep (lower is better)



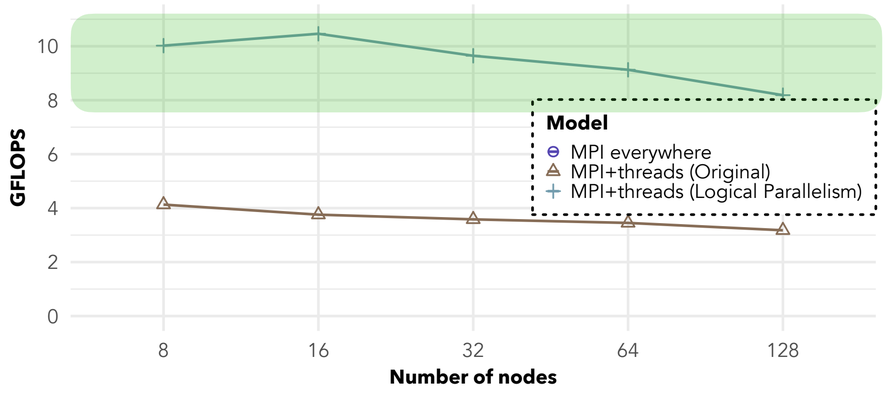
WOMBAT (Astrophysics)

Scientific throughput on 27 nodes (higher is better)



Circuit Legion app (Task-based runtime)

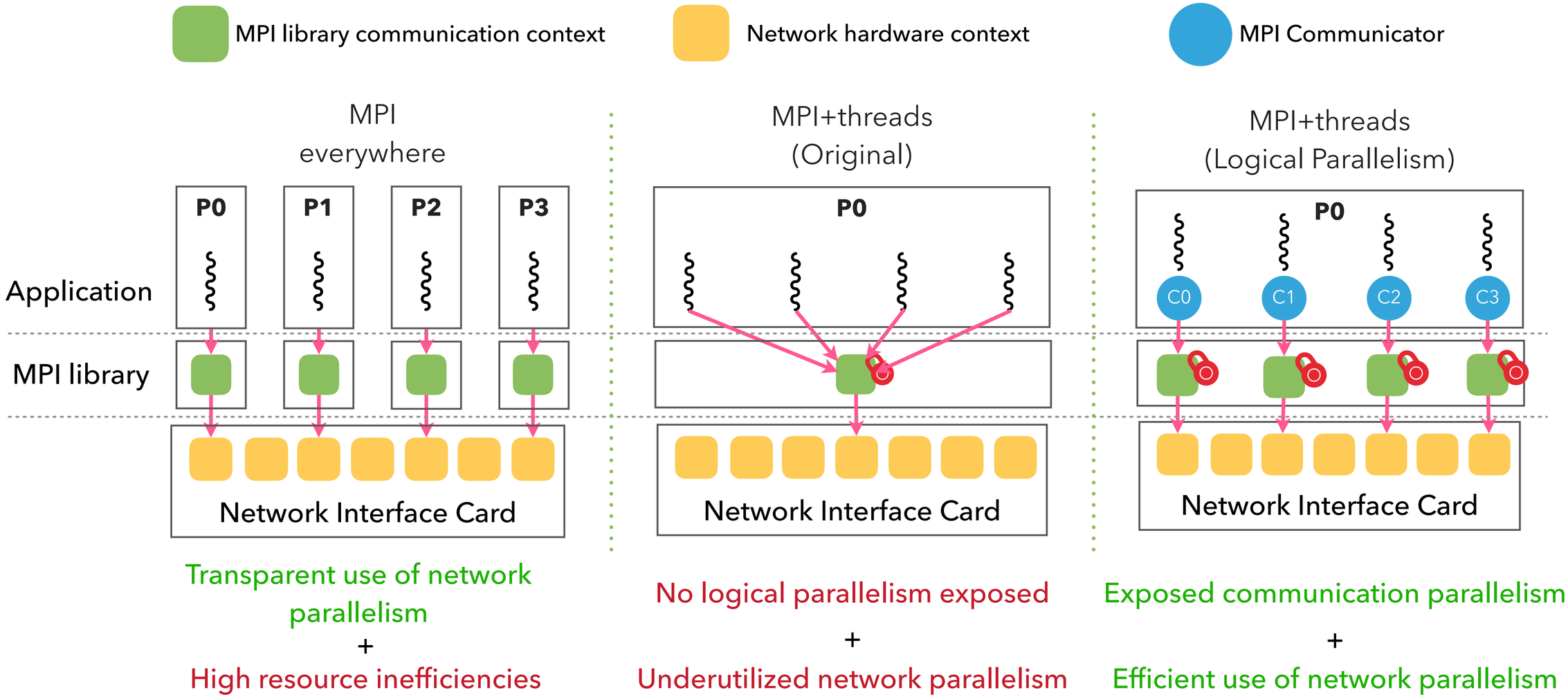
Throughput (higher is better)



- MPI+threads able to scale on modern architectures
- MPI+threads poses many challenges
- Efficient multithreaded communication overcomes the primary challenge

Logically parallel communication

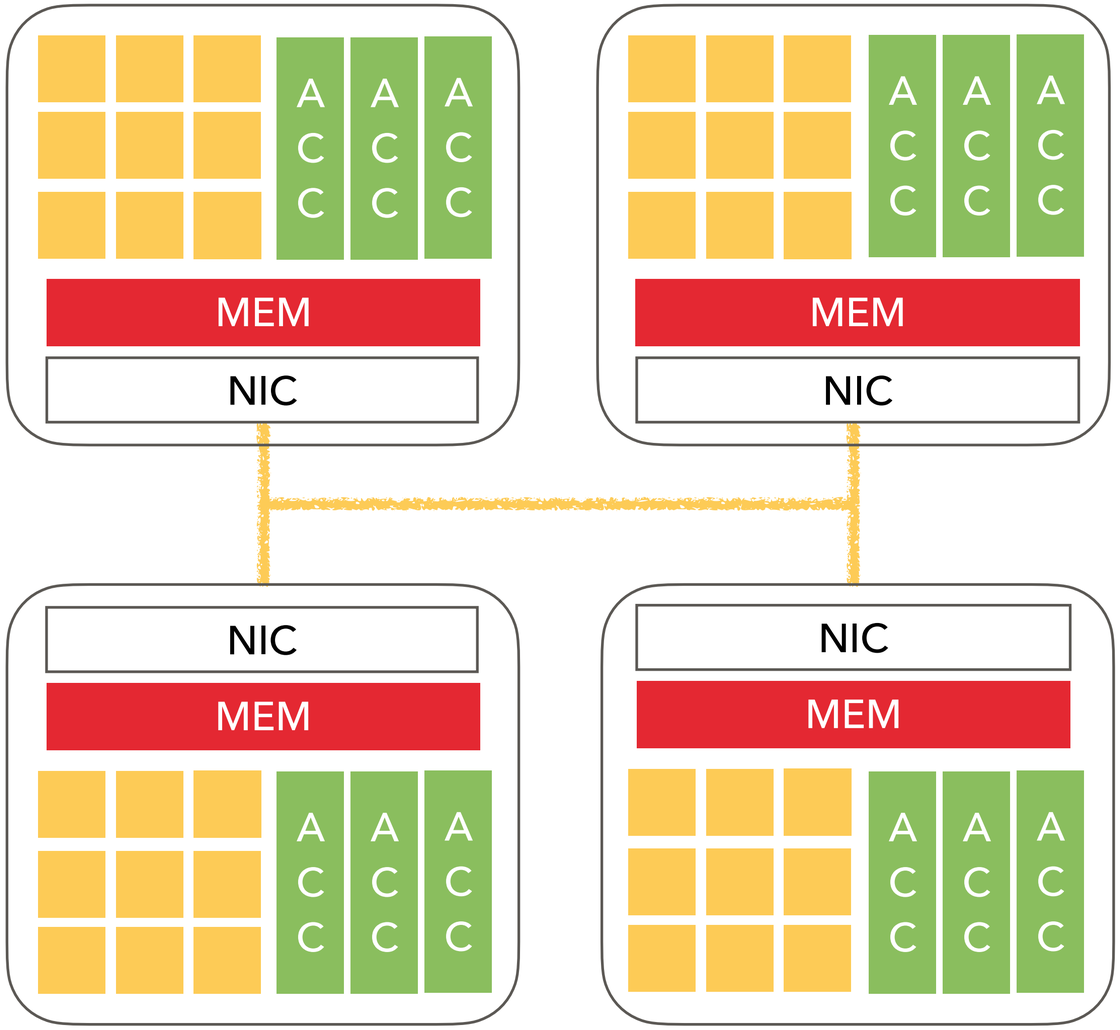
New capabilities of MPI libraries + New MPI programming features

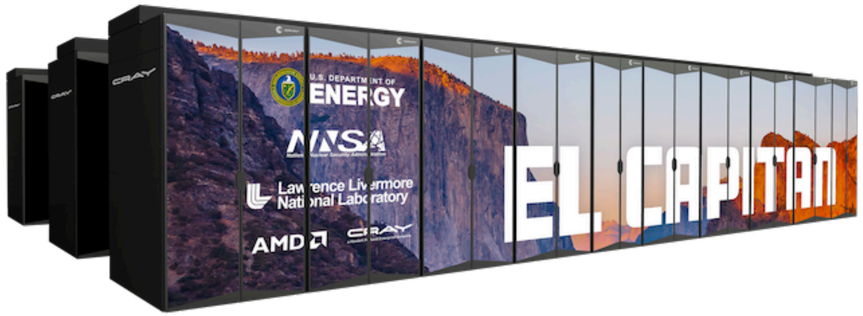


Process-based parallelism not an efficient way to scale apps

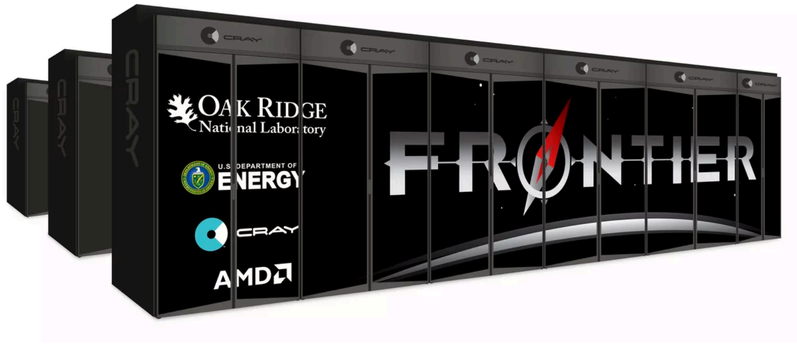
Interoperability between programming models critical for performance

- *Trend 2*: Increase in accelerators per node





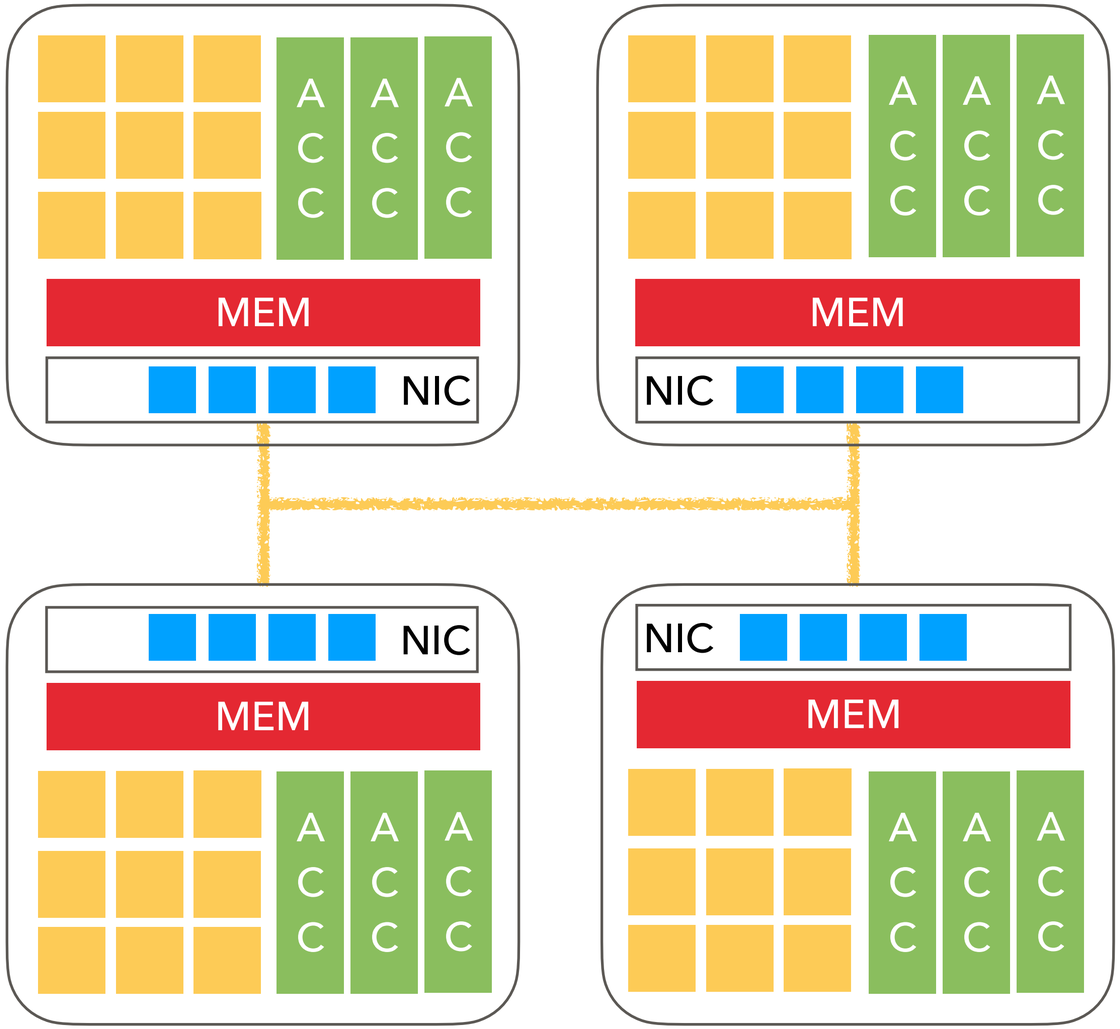
El Capitan node:
4 AMD GPUs

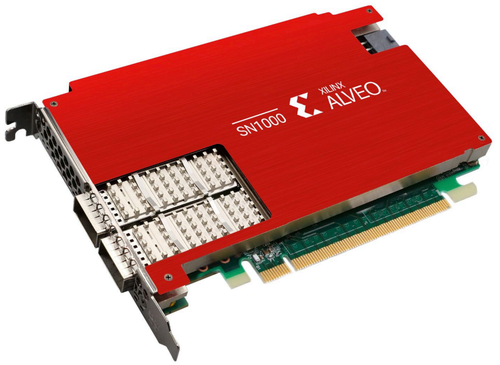


Frontier node:
4 AMD GPUs

Aurora node:
6 Intel GPUs

- *Trend 3*: Smart NICs per Node





Heterogeneous networking

- Accelerator-initiated communication
  - Effective in preventing kernel synchronization overheads
- In-network computing to offload communication tasks
  - Latency wins look slim with current architectures
  - How much and when to aggregate?
  - What is the best place for “smartness” of NICs?