In Network Compute

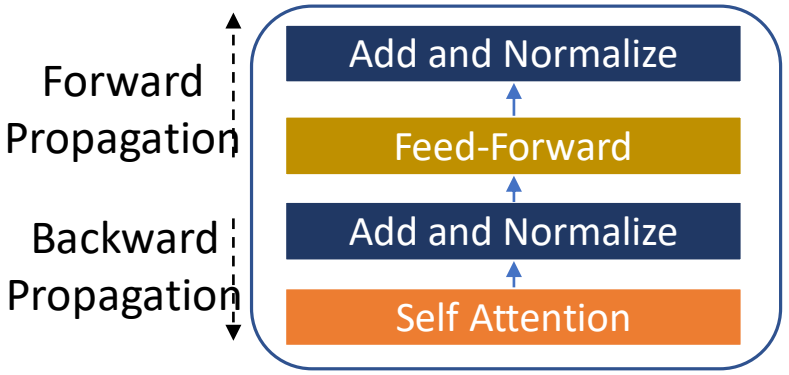
Surendra Anubolu

Broadcom Inc

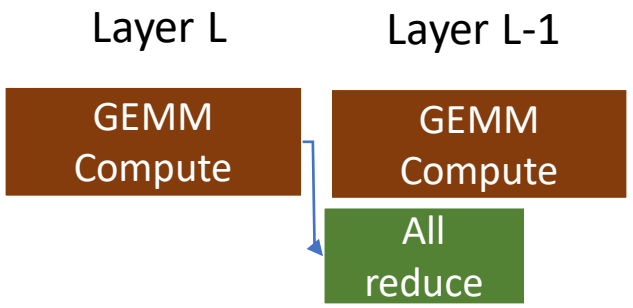
11/8/2023

Importance of All Reduce in Generative AI

- Training uses data parallel technique for scaling
- LLM training employs Tensor Parallelism (TP) to increase memory capacity
- Fabric is not scaling at the same rate as compute
  - Fabric is increasingly the limiter in the performance: Amdahl's Law



Transformer Layer



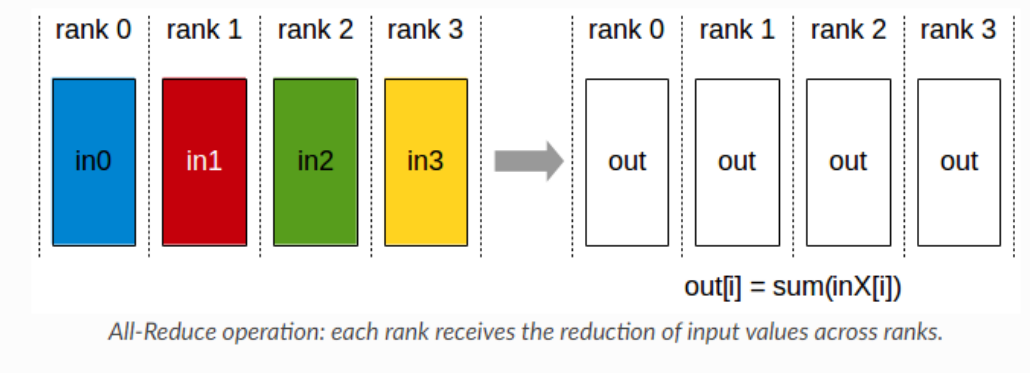
All Reduce - Data Parallel

All Reduce - Tensor Parallel

Compute vs communication

All Reduce operation – An example

- GPU nodes communicate over the network
- Data is exchanged and GPU servers do the addit
- When all the nodes sum up, the data out = sum(in)
- Reduced value (sum) is broadcasted down the tree
- At the end all the nodes have the same parameters



All-Reduce operation: each rank receives the reduction of input values across ranks.

- Next run starts

Observation

- Data traverses at least twice on the network
- Takes many steps to converge to the average (log(n))

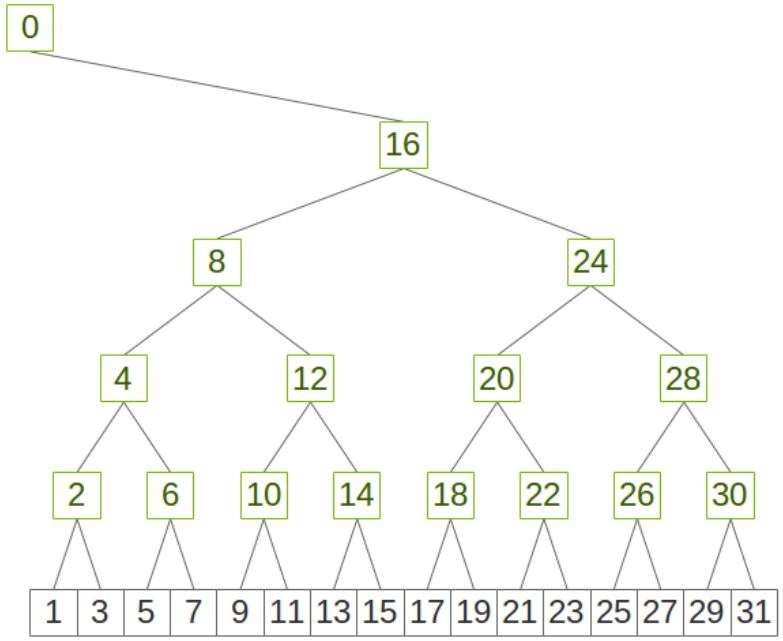
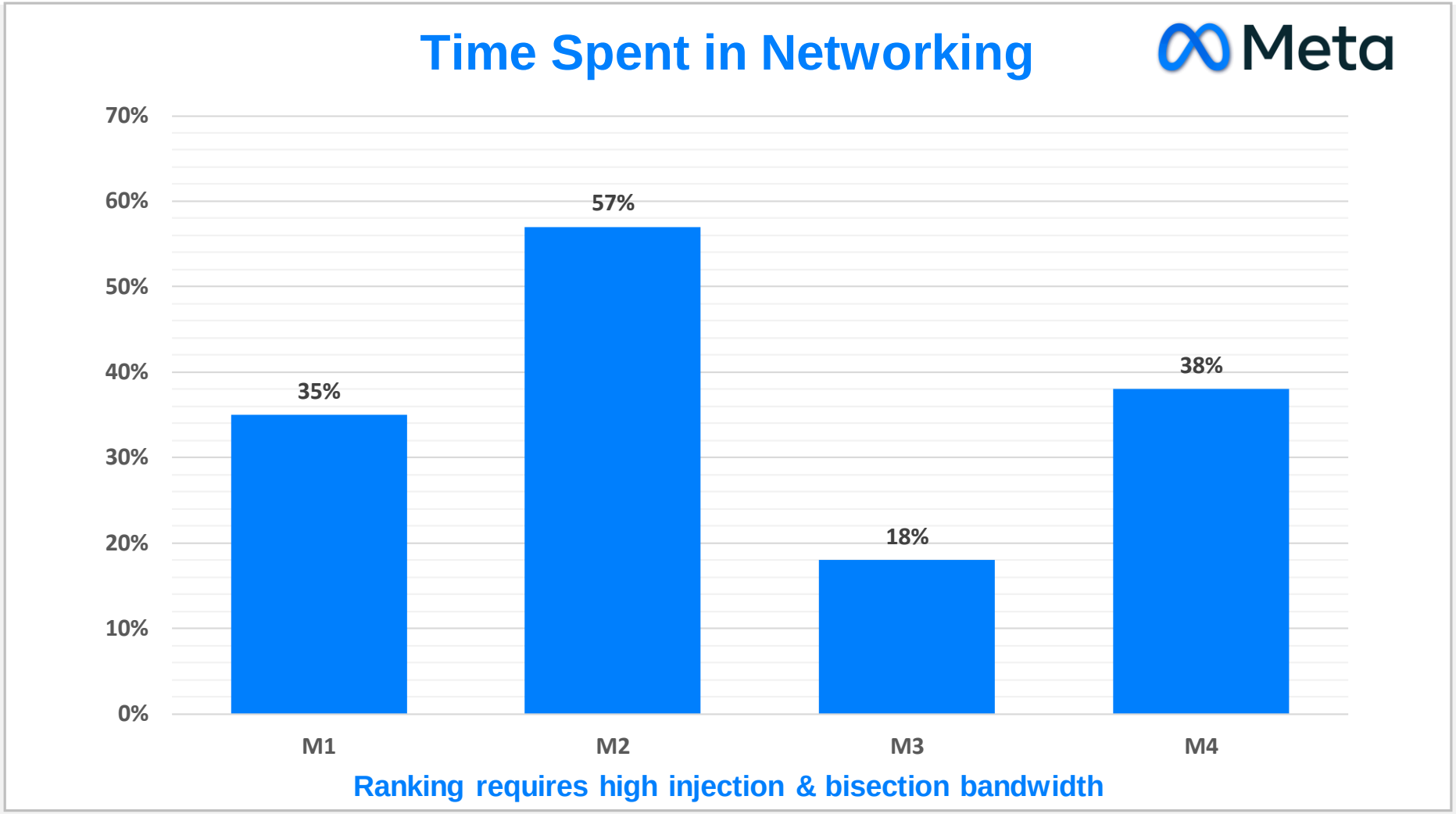


Figure 1. Binary tree using a power-of-two pattern

Each square represents a GPU node

Network I/O is Key For AI Workloads



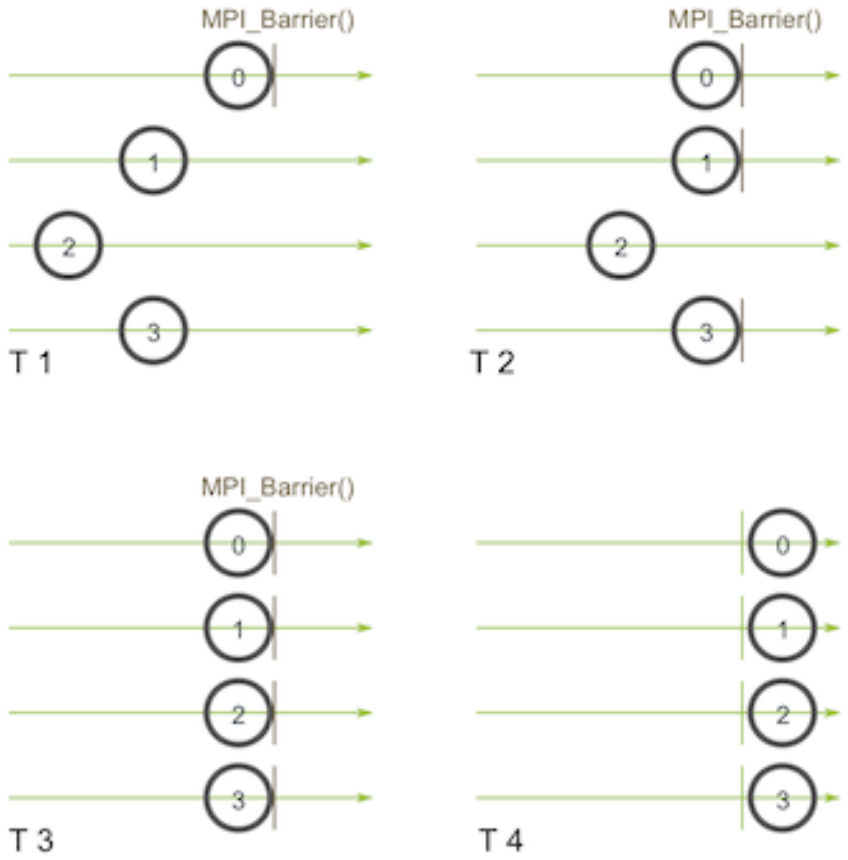
OCP keynote by Alexis Bjorlin at 2022 OCP Global Summit

M# = ML model #

Collectives

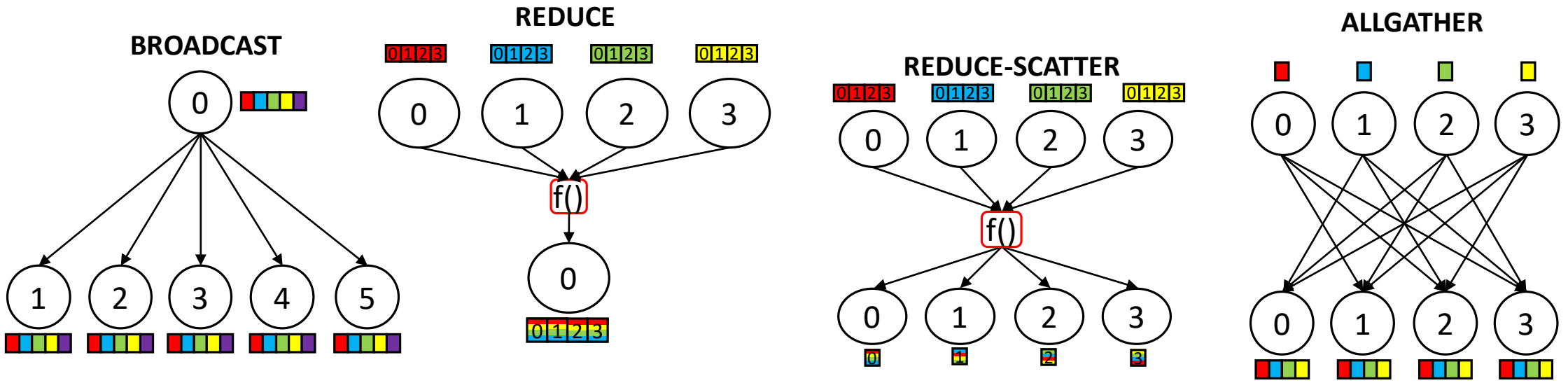
Common Collectives

- Barrier
- (All) Reduce
- Reduce Scatter
- All to All
- Broadcast
- All Gather



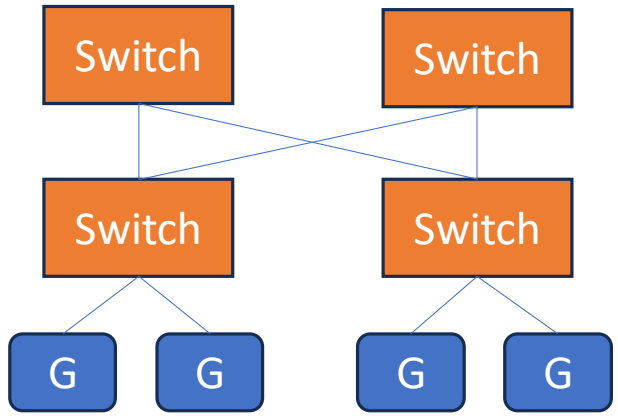
Example: Barrier

Collective Examples



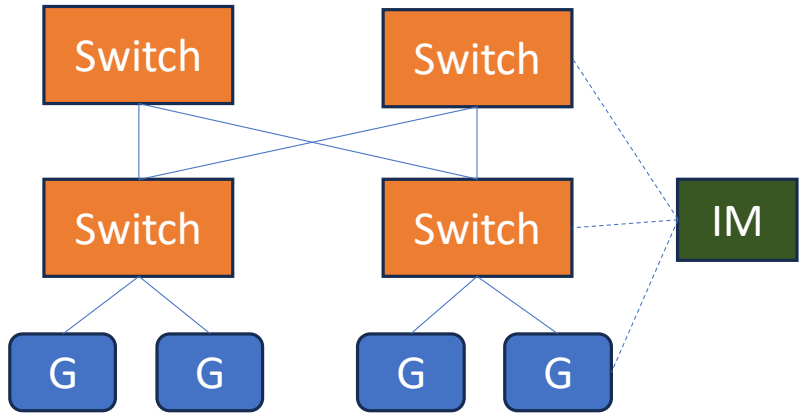
In Network Compute – Offload to the fabric

- Switches are natural choice for collective offload
- All the data passes through the switch
- Reduce the link load
  - Better performance
  - Lower power – potential oversub at higher layers
- Switches have high fanin and fanout
  - Large radix to 51T
- GPUs have limited bandwidth compared to network switch elements



Deploying In Network Compute

- Switch needs to be aware of the collectives
- Requires controller “INC Manager”
- INC Manager responsible for
  - Discovery of fabric elements
  - Process requests from GPU
  - Allocation of resources
  - Error reporting and tear down



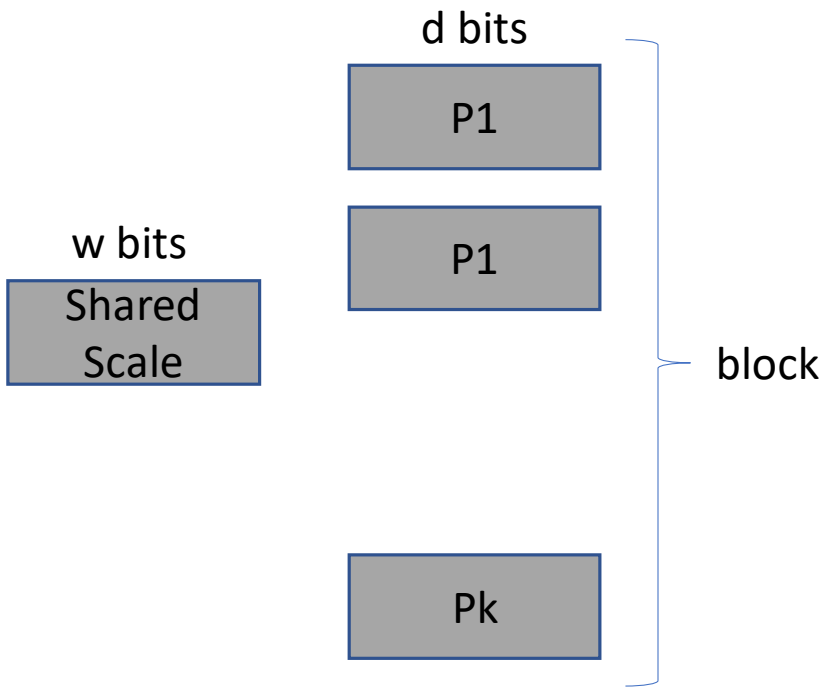
Number Formats

- INT32, INT8
- All the floating point numbers have
  - A sign bit
  - Range – exponent
  - Precision – mantissa
- New Number formats being discussed

Format	Exponent	Mantissa
FP32	8	23
FP16	5	10
BF16	8	7
FP8-E4M3	4	3
FP8-E5M2	5	2

Shared scale formats

- Scaling block size
  - Typical 32
- MXFP8, MXFP6, MXFP4, MXINT8
- Complexity for network based reduction



Implementation considerations

- Switch is the endpoint for RDMA transfers
- Block number formats
- Lightweight implementations
  - Driven by high radix and large throughput

- Interop between providers
- Error recovery and reporting
  - Large training clusters
  - Long job runs

INC Summary

- AI workloads spend large amounts of time on the collectives
- Fabric becomes part of the compute
  - Challenge: Keep up with number formats
- Power and cost savings
- Significant speed up with collective offload to the Fabric