Domain-Specific Interconnection Networks in the Era of Domain-Specific AI Supercomputer

John Kim (金東駿)
School of Electrical Engineering
KAIST
jjk12@kaist.edu

Work in collaboration with Hans Kasan, Hyojun Cho, Yassawe, Dennis Abts (NVIDIA)
An architect’s perspective for networking community

- Computer architecture is driven by technology (e.g., Moore’s Law) & application
  - Corollary 1: Interconnection networks are also driven by technology.
  - Corollary 2: Technology continue to evolve (e.g., optics, chiplets, etc.)
An architect’s perspective for networking community

• Computer architecture is driven by technology (e.g., Moore’s Law) & application
  ➔ Corollary 1: Interconnection networks are also driven by technology.
  ➔ Corollary 2: Technology continue to evolve (e.g., optics, chiplets, etc.)

• Communication performance is important but overall performance is more important.
  ➔ Corollary 1: There are many co-design opportunities.
  ➔ Corollary 2: Exposed communication is what matters.
An architect’s perspective for networking community

- Computer architecture is driven by technology (e.g., Moore’s Law) & application
  ➔ Corollary 1: Interconnection networks are also driven by technology.
  ➔ Corollary 2: Technology continue to evolve (e.g., optics, chiplets, etc.)

- Communication performance is important but overall performance is more important.
  ➔ Corollary 1: There are many co-design opportunities.
  ➔ Corollary 2: Exposed communication is what matters.

- In the end, it is all about perf/$ (or perf/W) [trade-off]
  ➔ Corollary 1: Cost-efficiency does matter. (e.g., Why did Google build TPU?)
  ➔ Corollary 2: Domain-specific architecture makes sense.
Submit relevant work to HPCA’24 😊

The International Symposium on High-Performance Computer Architecture (HPCA) is the premier forum for new ideas and research results in computer architecture. In 2024, the 30th edition of HPCA will be held in Edinburgh, Scotland, UK.

HPCA-30 will be held in conjunction with PPoPP, CGO and CC.

View the Calls for Papers: Regular Track CFP · Industry Track CFP
Challenges in Life
Challenges in Life
Challenges in Life

Unpredictability
Unpredictability

Congestion

Tail latency

Reliability
Today’s Talk

• Introduction/Background

• High-Performance Interconnection Networks

• Case for Domain-specific Network

• Case I: Groq *software-managed* scale-out network

• Case II: Collective-Communication (AllReduce)

• Summary
Who builds “supercomputer”? 

“World’s First Deep Learning Supercomputer”

Cray

Microsoft

NVIDIA

Tesla
Computer Systems & Architecture

Interconnect: Moving bits around
Interconnection Networks

- Microarchitecture - router organization
- Topology - “roadmap” of the network
- Routing - which path a packet takes
- Flow Control - allocation of network resources
Exponential growth in computing resources

Larger Models
More memory
More Accelerators

https://huggingface.co/blog/large-language-models
Implications on Computer Systems & Architecture

Zeta-scale of FLOPS

10s of TBs

10s of GBs

Processor

Computer Architecture

10s of TBs

10s of GBs
Domain-Specific (AI) Supercomputer

A Domain-Specific Supercomputer for Training Deep Neural Networks

Google’s TPU supercomputers train deep neural networks 50x faster than general-purpose supercomputers running a high-performance computing benchmark.

BY NORMAN P. JOUPPI, DOE HYUN YOON, GEORGE KURIAN, SHENG LI, NISHANT PATIL, JAMES LAUDON, CLIFF YOUNG, AND DAVID PATTERSON
Domain-Specific (AI) Supercomputer

- TPU v4 (Google)
- Dojo (Tesla)
- ZionEX (Meta)
- DGX H100 & Superpod (NVidia)
- WSE 2 & CS-2 (Cerebras)
A New Golden Age for Computer Architecture:

Domain-Specific Hardware/Software Co-Design, Enhanced Security, Open Instruction Sets, and Agile Chip Development

John Hennessy and David Patterson
June 4, 2018

- HW-centric
  - Only path left is Domain Specific Architectures
  - Just do a few tasks, but extremely well
# Domain-Specific Memory(?)

### Processing In Memory

<table>
<thead>
<tr>
<th>Memory-centric Computing with SK Hynix’s Domain-Specific Memory</th>
<th>Yongkee Kwon, SK Hynix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Samsung AI-cluster system with HBM-PIM and CXL-based Processing-near-Memory for transformer-based LLMs</td>
<td>Jin Hyun Kim, Samsung</td>
</tr>
</tbody>
</table>

---

**Hot Chips 2023 Program**

- SK Hynix AiM
- Samsung HBM-PIM
- Samsung AXDIMM
Why not domain-specific *network*?
Today’s Talk

• Introduction/Background

• High-Performance Interconnection Networks

• Case for Domain-specific Network

• Case I: Groq *software-managed* scale-out network

• Case II: Collective-Communication (AllReduce)

• Summary
Bandwidth growth

[data from Nic McDonald]
High-performance Computing Interconnect

YARC ROUTER [ISCA’06]

Flattened Butterfly [ISCA’07, MICRO’07]

Dragonfly [ISCA’08]

Figure 4 DGX-1 uses an 8-GPU hybrid cube-mesh interconnection network topology. The corners of the mesh-connected faces of the cube are connected to the PCIe tree network, which also connects to the CPUs and NICs. The cube-mesh topology provides the highest bandwidth of any 8-GPU NVLink topology for multiple collective communications primitives, including broadcast, gather, and all-gather, which are important to deep learning. Using NVLink connections to span the gap between the two clusters of four GPUs relieves pressure on the PCIe bus and on the inter-CPU SMP link, and avoids staging transfers through system memory when transferring across the two clusters.
Challenges : Adaptive Routing
Challenges: Handling Congestion

**Network congestion**
- Congestion within network
- Adaptive routing can help

**Endpoint congestion**
- Caused by the endpoints
- Network routing cannot help (can make it worse)
Domain-Specific Architectures

[Microsoft]
Domain-Specific Networks

General-Purpose Networks

Flexibility

- Ethernet
- Infiniband (IB)

Cray/HPE Slingshot

NVIDIA NVLink/NVSwitch

Domain-Specific Networks

Efficiency (perf/$)

- IBM BlueGene/Q
- NVIDIA NVLink/NVSwitch
- Intel/Google TPU
- DE Shaw Anton
- Tesla Dojo
- Groq Scale-out TSP

ASIC
Domain-Specific Networks Characteristics

- Latency & Bandwidth
- Custom Communication Protocol
- Exploit (known) communication patterns
- Low Cost & Complexity

[Abts & Kim HOTI’23]
Today’s Talk

• Introduction/Background

• High-Performance Interconnection Networks

• Case for Domain-specific Network

• Case I: Groq *Software-Scheduled* Scale-out Network [ISCA’22]

• Case II: Collective-Communication (AllReduce)

• Summary
Challenges in Interconnect

Challenges: Adaptive Routing

Challenges: Handling Congestion

Network congestion
- Congestion within network
- Adaptive routing can help

Endpoint congestion
- Caused by the endpoints
- Network routing cannot help (can make it worse)
Domain-Specific Networks Characteristics

- Latency & Bandwidth
- Custom Communication Protocol
- Exploit (known) communication patterns
- Low Cost & Complexity

[Abts & Kim HOTI’23]
Domain-Specific Networks Characteristics

- Latency & Bandwidth
- Custom Communication Protocol

Route Tensors, Not Packets
320B vector packet format
~2.5% packet overhead
Domain-Specific Networks Characteristics

- Latency & Bandwidth
- Custom Communication Protocol
- Exploit (known) communication patterns ➔ Software scheduled network
Domain-Specific Networks Characteristics

- Latency & Bandwidth
- Custom Communication Protocol
- Exploit (known) communication patterns
- Low Cost & Complexity

No "switches"
Nodes == switches
Groq Tensor Streaming Processor (TSP)

220MiB SRAM

[Abts et al ISCA'20]
Software-scheduled (Deterministic) Network

Conventional Scale-out Network

- Per-hop hardware router arbitration
- Hardware-based global adaptive routing
- Congestion sensing in the network through backpressure

Interconnection network
Scale-out Organization

INDIRECT NETWORK

DIRECT NETWORK

SOFTWARE-SCHEDULED DIRECT NETWORK
Low-Diameter Network

- The total observed latency and variance increases with the number of hops in the network.
  - Dragonfly is a hierarchical topology that minimizes the number of hops taken
    - Local group topology
    - All-to-all global topology
    - Exploits packaging locality
- Local topology 2x speedup
- Scalability to 10s-of-thousands of TSPs
Deterministic Adaptive Routing

Conventional Network

- Commonly done based on network backpressure
- Reactive approach makes the routing decision difficult, increases latency, and increases hardware complexity
- Network latency is unpredictable

Software-scheduled Network

- Avoids congestion
- Enables maintaining a deterministic TSP architecture to scale to a multi-node deterministic network execution
Load Balancing the Physical Links

Always taking advantage of non-minimal is not beneficial

- Additional hop increases latency

Need to determine how many non-minimal routes to consider

Rule of thumb

- Small tensors ⇒ only use minimal path to minimize latency
- Large tensors ⇒ exploit non-minimal paths to maximize bandwidth

Benefit of non-minimal routing is a function of

- Per-link (channel) bandwidth
- Per-hop latency
- Message size
AllReduce Comparison Results

- Only a handful of cycles to Read(vector) → Send(vector) enables fine-grained communication across the 16 directly connected links on each TSP
- Comparison made with 8 GPU A100 system with NCCL
- A100 system has approximately 3x higher network channel bandwidth
- When normalized, Groq TSP matches the bandwidth at large tenor size while significantly improving bandwidth at intermediate tensor size
Today’s Talk

• Introduction/Background

• High-Performance Interconnection Networks

• Case for Domain-specific Network

• Case I: Groq Software-Scheduled Scale-out Network [ISCA’22]

• Case II: Collective-Communication (AllReduce)

• Summary
Challenges in Scaling Systems

Reliability!!!
Can we potentially live with unreliable networks?

- Not all gradients are needed.
- What happens if we drop gradients?
- Prior work: Gradient Sparsification
  - Overhead?
  - Can we just randomly drop gradients?
Impact of randomly dropping gradients
Baseline AllReduce

Node 1

Node 2

Node 4

Node 3
Skip AllReduce

Node 1

Node 2

Node 4

Node 3
Preliminary Performance Results

![Graph showing test accuracy and normalized iteration time for different algorithms.]

- **Test Accuracy (%)**
  - Baseline
  - Skip 1RS
  - Skip 2RS
  - Skip 1RS1AG
  - Progressive

- **Normalized Iteration Time**
  - Baseline
  - Skip 2RS
  - Skip 4RS
  - Skip 2RS2AG
  - Progressive
Domain-Specific Network for Deep Learning

• Can we sacrifice reliability by exploiting the characteristics of Deep Learning communication? (e.g., AllReduce)

• What are the hardware implications? Significant reduction in complexity?

• Can we enable more scalable systems?

• What about other type of communication patterns? (e.g., All-to-All)
Summary

• Increasing demand for compute (and memory) results in the need for AI supercomputer.

• Communication or movement of data (interconnection network) is a key challenge to enable (efficient) scalable systems.

• Case for Domain-Specific Networks
  – Exploit (known) communication pattern
  – Minimize communication cost
  – Maximize impact on overall system performance

• Need to maximize utilization of interconnect channel bandwidth
Domain-Specific Interconnection Networks in the Era of Domain-Specific AI Supercomputer

John Kim (金東駿)
School of Electrical Engineering
KAIST
jjk12@kaist.edu
Physical Topology in Networks

- Ring
- Mesh
- Torus
- Tree
**Logical Topology for AllReduce**

**Ring**
- Node A
  - Node B
  - Node D
- Node C

**Tree**
- Node A
  - Node B
    - Node C
    - Node D

1 2 3 4
Overlap communication with computation

"Out-of-order"

"In-order"
Physical Topology – detour routes

• Detour routes == non-minimal routing

• Pre-defined (or deterministic) non-minimal routes based on the communication pattern.

• Introduces some (small) overhead but maximizes communication performance
Physical Topology

- 8-node DGX system
- Cube-mesh topology
Gradient Distribution

ResNet-50 on CIFAR-10

BERT on SQuAD
Performance with Random Dropping

ResNet-50 on CIFAR-10

BERT on SQuAD
Overhead of top-k Search

Overhead of searching top-10% of the gradients
Performance with SkipReduce

ResNet-50 on CIFAR-10

BERT on SQuAD
What is an Interconnection Network

• “A programmable system that enables fast data communication between components of a digital system.”
[Dally & Towles]
  – Sharing of expensive communication resources
  – Provide a structured way to organize communication
Interconnection Networks

Supercomputers

Processor Interconnect

NoC

CPU/GPU

Accelerators

I/O systems

Infiniband

Interconnection networks

Cray XC50

AI Supercomputer
What is an Interconnection Network

• “A programmable system that enables fast data communication between components of a digital system.” [Dally & Towles]
  – Sharing of expensive communication resources
  – Provide a structured way to organize communication

• Four components of interconnection networks
  – Microarchitecture - router organization
  – Topology - “roadmap” of the network
  – Routing - which path a packet takes
  – Flow Control - allocation of network resources
Training Time with SkipReduce

ResNet-50 on CIFAR-10

BERT on SQuAD