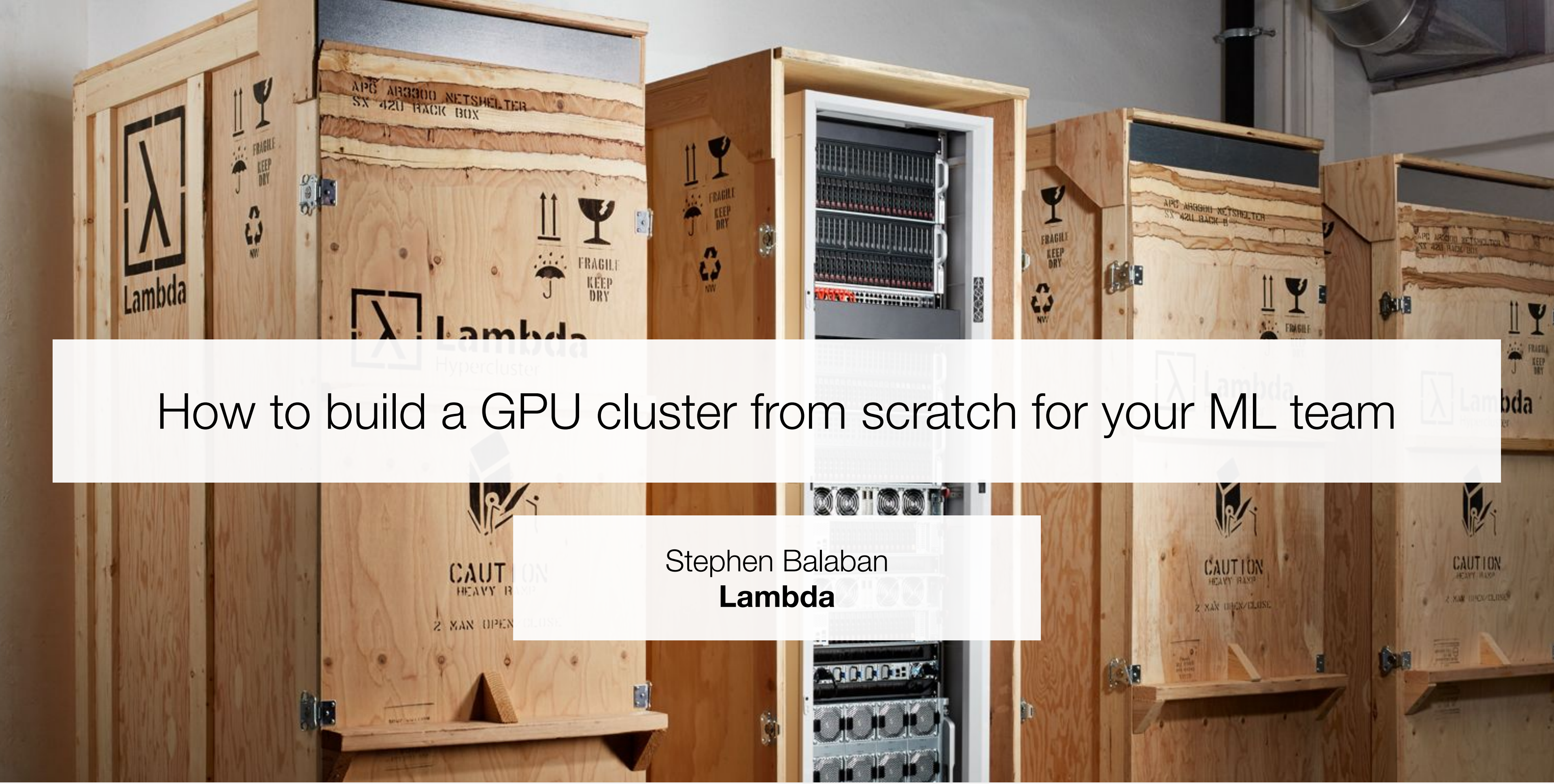


How to build a GPU cluster from scratch for your ML team

Stephen Balaban
Lambda

Who am I?

I’m the co-founder and CEO of **Lambda**. We’ve built large scale GPU clusters for the Fortune 500, the world’s leading academic research institutions, and the DOD.

I’m also the lead architect of the **Lambda Echelon**, a turn-key GPU cluster. This talk is based on the Lambda Echelon reference design whitepaper and the experience we’ve gained deploying these large clusters.

Overview

- Introduction
- Cluster use case overview
- Cluster design
    - Compute
    - Storage
    - Cluster Networking
    - Data center floor planning & power distribution checklist
    - Software
- Rack design
    - Rack elevations
    - Power basics
    - Rack power distribution
- Node design
    - GPU selection
    - Node PCIe topology / NVLink topology / real life examples
    - GPUDirect RDMA
- Putting it all together
- Citations

5 Stages of GPU Cloud Grief

It all starts with the **shock** of an expensive cloud bill.

```
Greetings from Public Cloud Provider,

This e-mail confirms that your latest billing statement, for the account ending in
1234, is available on the Cloud web site. Your account will be charged the
following:

Total:  $18,445.92

You can see a complete break down of all charges on the Billing & Cost Management
page.

Thank you for using the public cloud.

Sincerely,
Public Cloud Provider
```

Stage 1 - Denial

"This won't happen again next month."

Stage 2 - Anger

“The bill doubled again!”

Stage 3 - Bargaining with your account manager.

```
Hi Team,

Are you available in the next few days for a 15 min call to discuss lowing your
public cloud spend using reserved instances?

Based on your January usage, purchasing a 1-year reserved instance would save you
up to 20% on your public cloud bill!

Thank you for being a loyal public cloud customer.

Cheers,
Public Cloud Provider
```

Stage 4 - Depression

"Spot instances and reserved instances aren't enough, this is hopeless."

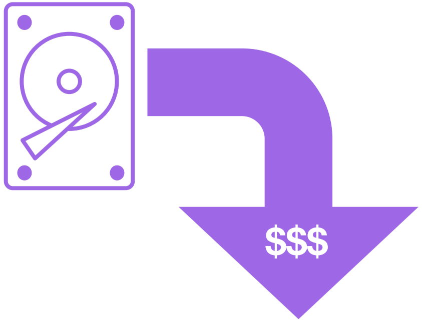
Stage 5 - Acceptance

"GPU cloud services are expensive. Managing hardware is scary."

Solution

Learn how to build a GPU cluster.

Good reasons to consider building an on-prem cluster



Big data sets & expensive egress

Data sovereignty & security

$$$

More compute for less money

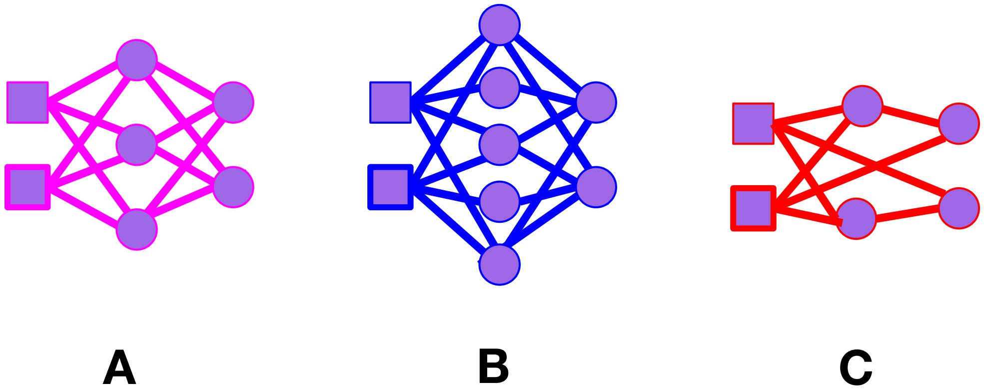
Rule Number 1:
Know your ML team.

Your use case should inform the cluster design

- **Hyperparameter search**
  - “Finding the best model.”
- **Large scale distributed training**
  - “Quickly training a model.”
- **Production inference**
  - “Deploying the model at scale to production.”

Hyperparameter search

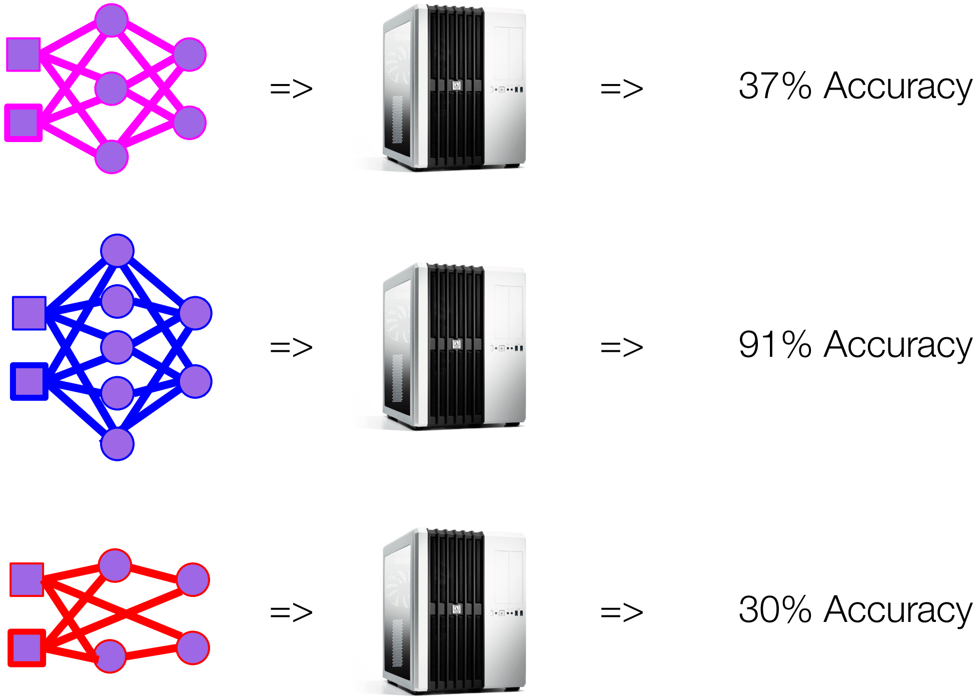
Which neural network performs best (highest accuracy)?



Hyperparameter Search

Answer: ____?_____

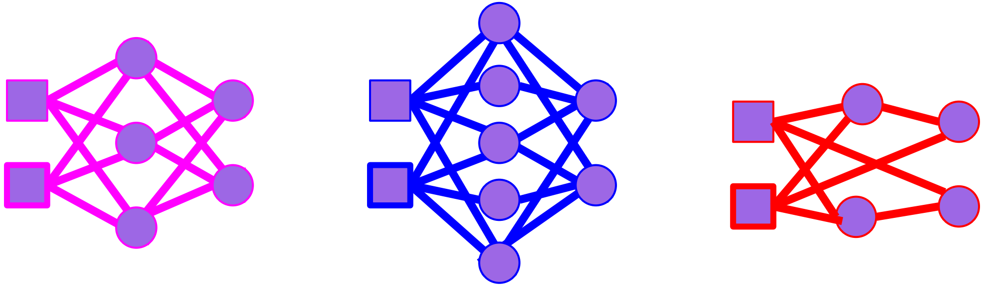
Hyperparameter search scoring



Hyperparameter Search

Hyperparameter search

Give a score to each, select the highest score.



A
Score: 37%

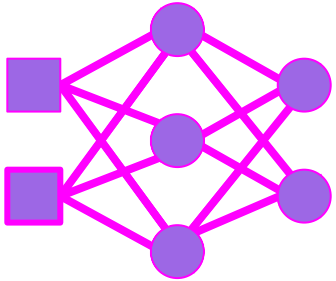
B
Score: 91%

C
Score: 30%

Hyperparameter Search

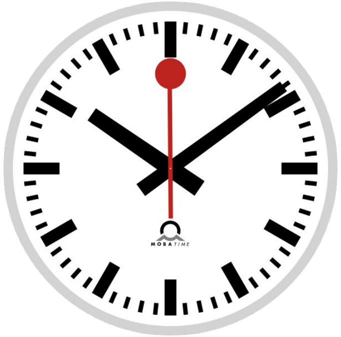
Answer: ____B_____

Each training run can take 1 to 48+ hours!



=>

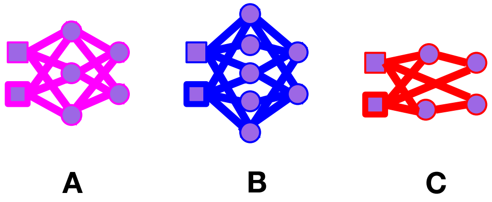
=



Hyperparameter Search

Most people start small

Which neural network performs best?



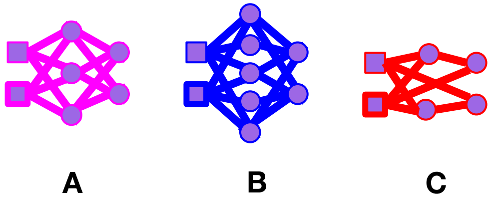
Answer: ____?____

=>

Hyperparameter Search

Then scale to faster GPUs

Which neural network performs best?



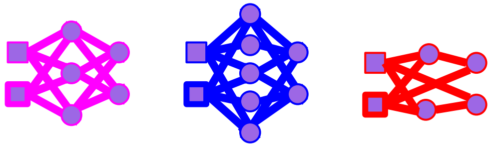
Answer: ____ ? _____

=>

Hyperparameter Search

Then scale to multiple servers to run the jobs simultaneously

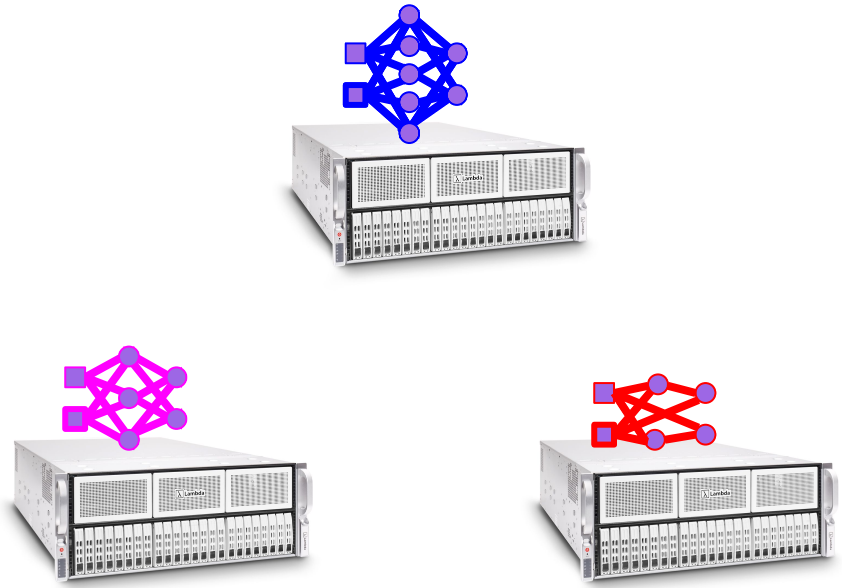
Which neural network performs best?



A **B** **C**

Answer: ____?____

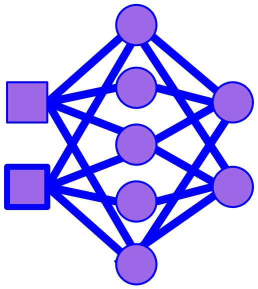
=>



Hyperparameter Search

Large scale distributed training

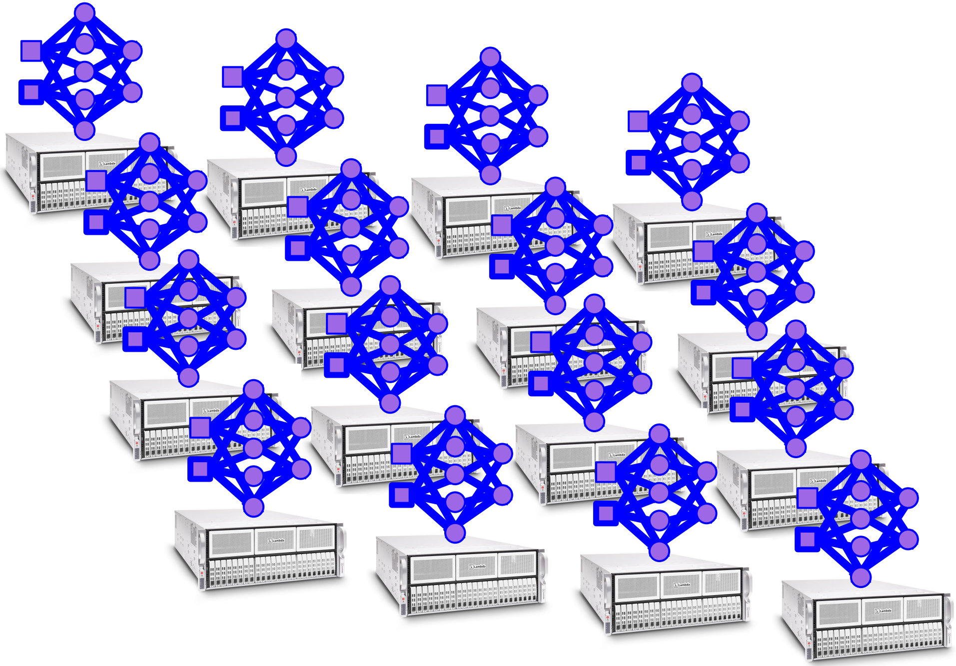
Train this model as fast as possible



B

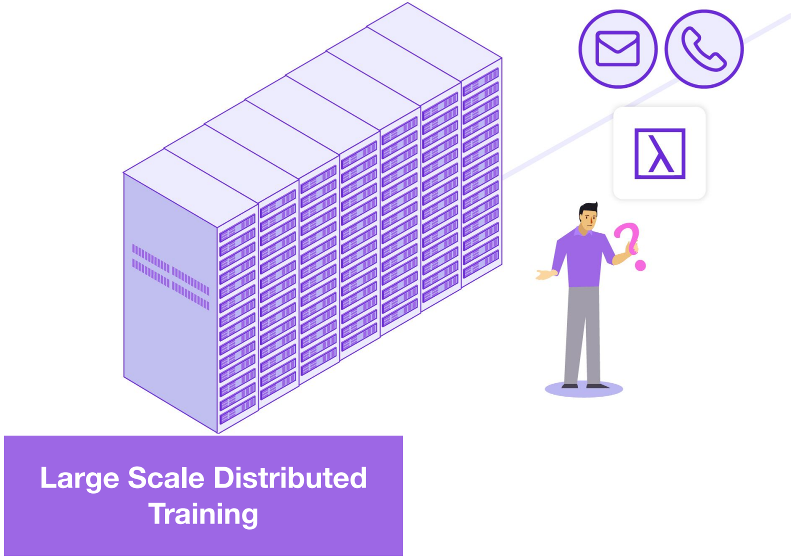
Large Scale Distributed Training

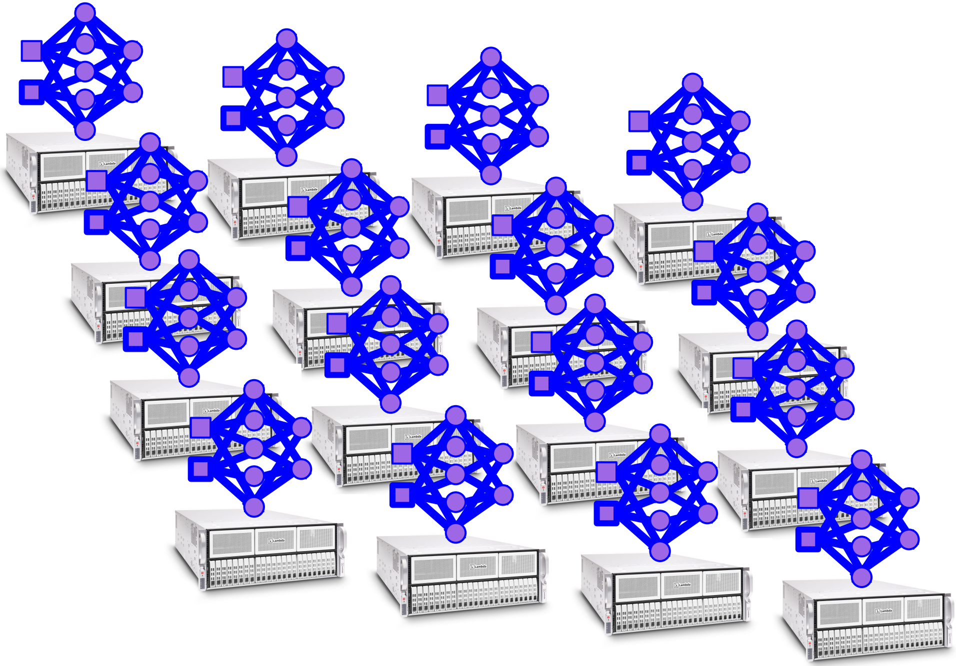
Have multiple servers help with the training



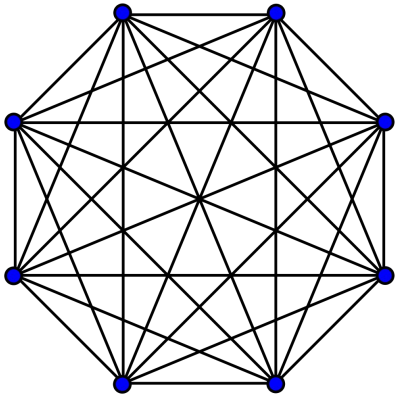
Large Scale Distributed Training

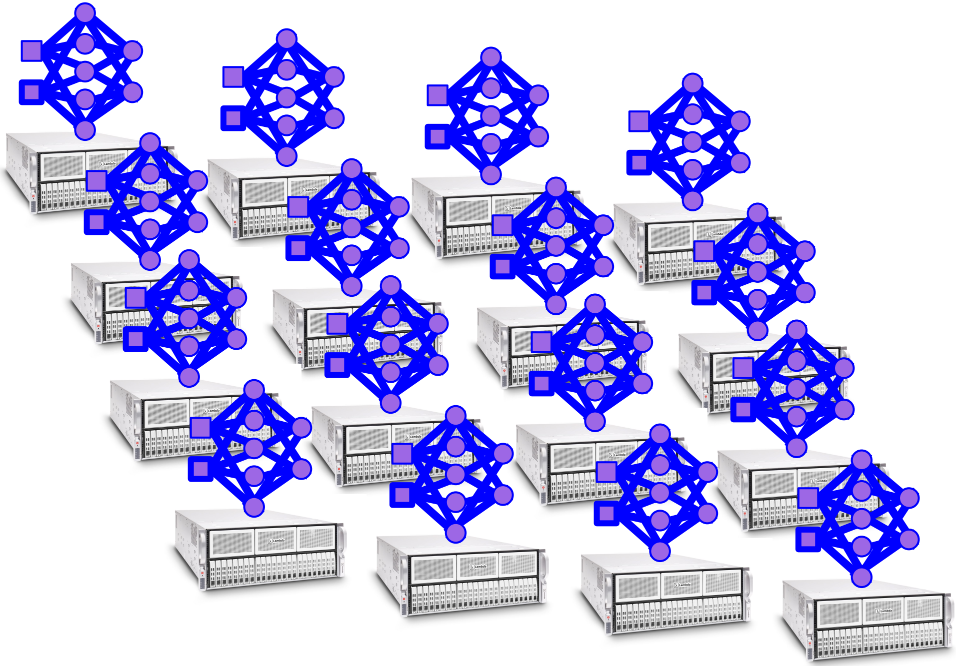
Coordinating the servers is hard





Lots of server-to-server communication = high speed network required

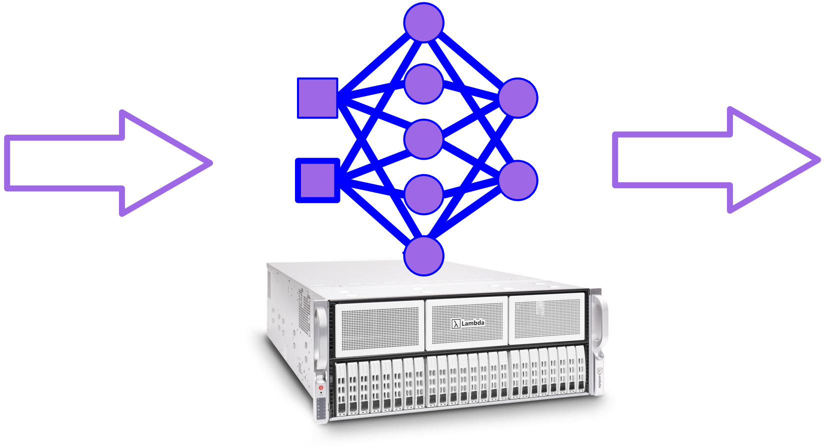




Large Scale Distributed Training

Production inference

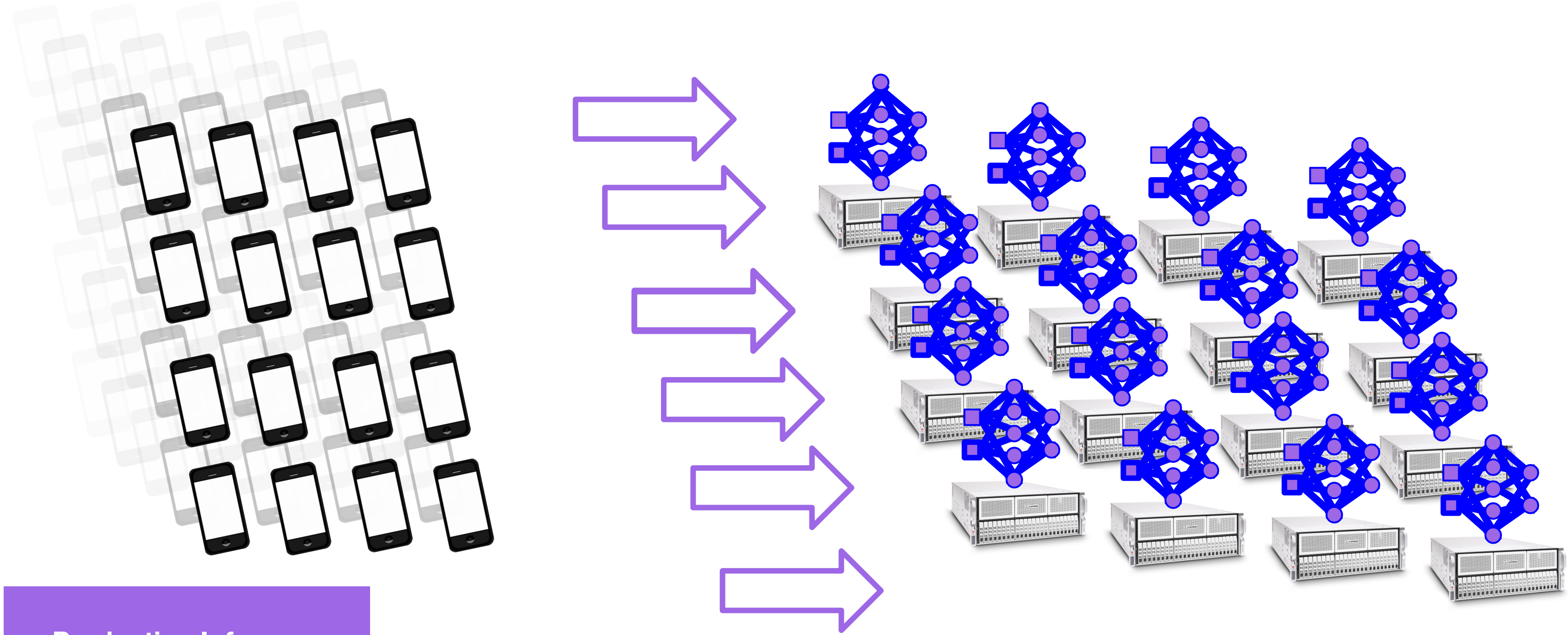
Hey Lambda, translate
"Hello" to Chinese



你好 - nǐ hǎo

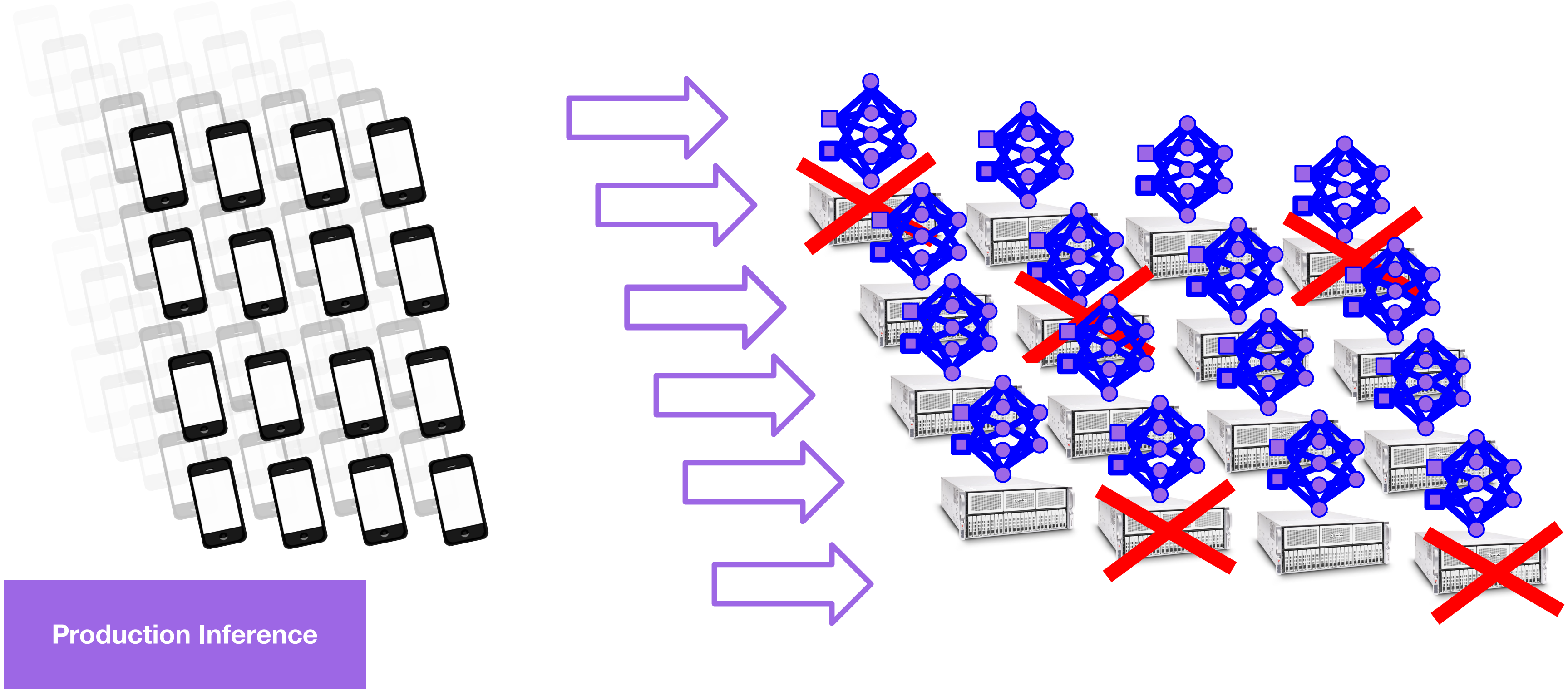
Production Inference

Inference clusters handle thousands of simultaneous requests



Production Inference

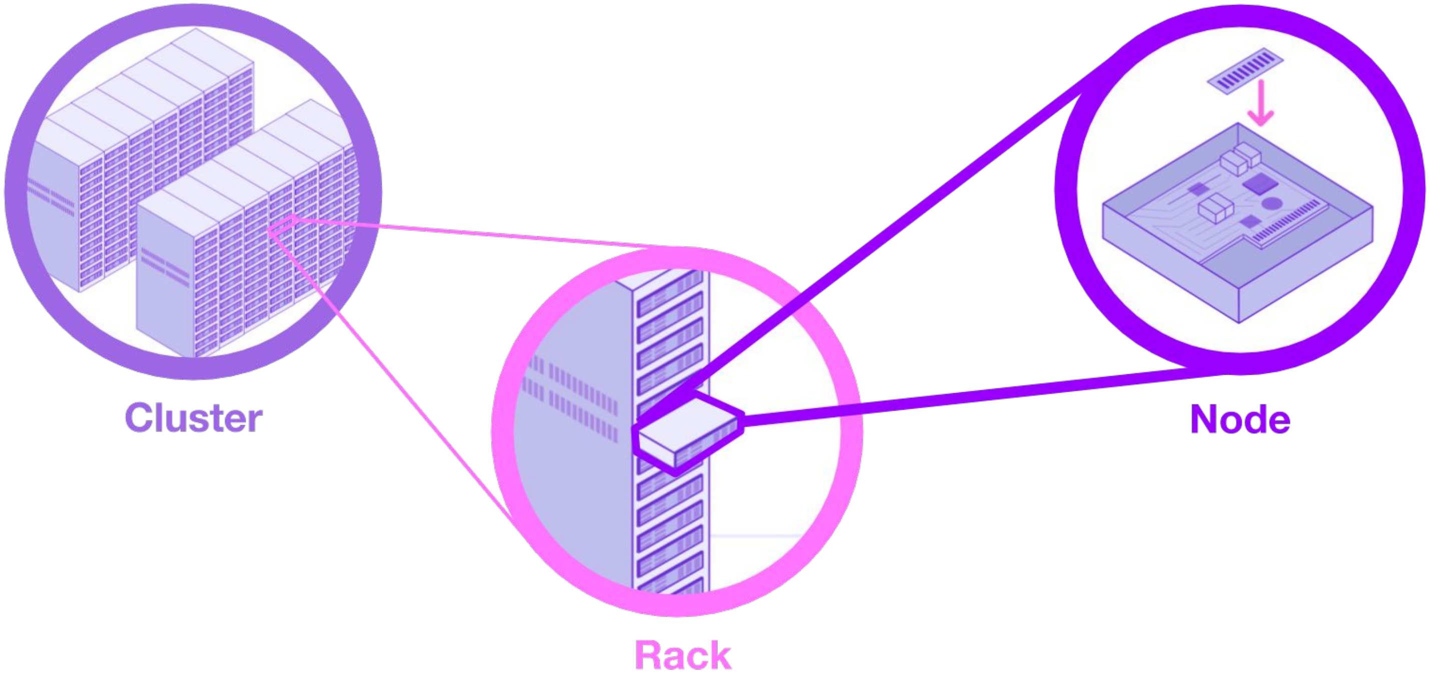
Inference clusters must be robust to outages



What's the right cluster for each use case?

	Hyperparameter search	Large scale distributed training
Node to Node Bandwidth	Low	High
Key Metric	Training throughput / $	Time to train a single large model
Price	$$	$$$$
1+1 Redundancy	Optional	Optional
Operating Mode	Usually offline, job-queue based	Usually offline, job-queue based
	Production inference	
Node to Node Bandwidth	Low	
Key Metric	High availability & throughput	
Price	$$$$$	
1+1 Redundancy	Critical	
Operating Mode	Often online, real time results	

Three levels of abstraction



The three levels of cluster design

1. **Cluster design:** the highest level of abstraction. Entire racks are just dots on the screen. Data center floor plans, capacity planning (power, network, storage, compute), and network topologies are the main output products at this level.

2. **Rack design:** Rack elevations describe the layout and exact position of individual nodes in a particular rack. Cable and port counts are important details at this layer.

3. **Node design:** Individual node bills of materials and cluster design goals drive component selection for each node.

Software: it's important to remember that there is a two way design dependency between the cluster hardware and the software that runs on it. Selecting the right software depends on the use case and scale of your planned deployment.

Work products for each layer

Abstraction Layer	Work Products	Goals
Cluster	• Cluster BOM (Bill of Materials) • Data center floor plan & checklist • Capacity plan • Network topology • Installation plan	• Working within constraints of data center floor space and power availability • Connecting all of the racks together • Determining the overall storage/compute/network capacity
Rack	• Rack BOM • Rack elevation	• Properly distributing the capacities planned for in the cluster stage • Rack level power density & capacity • Node placement within rack
Node	• Node BOM	• Hitting the performance requirements defined by the compute capacity of the cluster
Software	• Software architecture • Software installation plan	• Tying all of the hardware together so that it's easy to use and administer

Cluster design

Cluster architecture

Cluster architectures have five components:

1. **Compute:** to do the actual work. GPU & CPU nodes.
2. **Storage:** to serve data sets and store trained models / checkpoints.
3. **Networking:** multiple networks for compute, storage, in-band management, out-of-band management.
4. **Data center power distribution & floor planning:** understanding the electrical and physical layout of the deployment data center informs the rack elevations, cable lengths, and networking.
5. **Software:** cluster orchestration, job scheduling, resource allocation, container orchestration, and node level software stacks.

Compute

You need to answer three questions:

1. How much compute do you need?
2. How fast do you want the compute to talk to each other?
3. How fast do you need the compute to be able to talk to the storage?

Storage

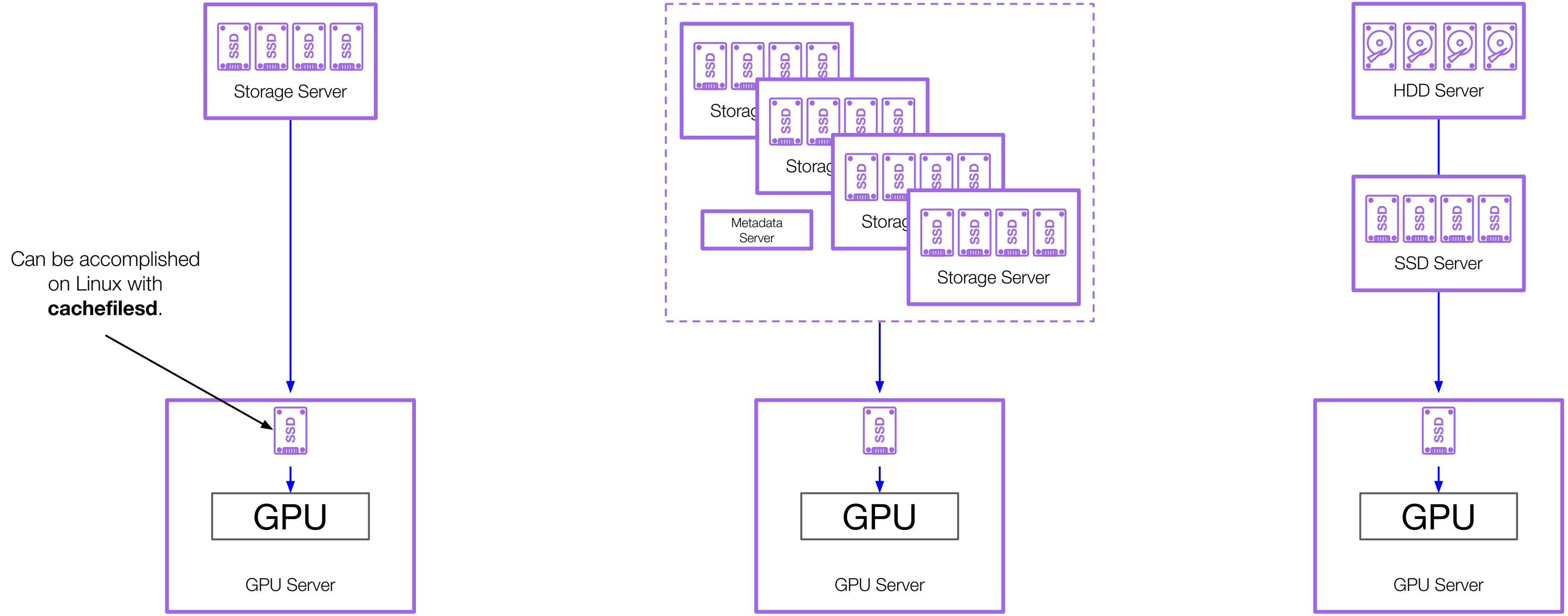
- Often, storage clusters will become the bottleneck in a highly optimized cluster. It's important to remove this bottleneck because it reduces the utilization of the most expensive part of the cluster, the compute.

- There are a two ways to design a storage cluster:

    - Work with a storage partner

    - Roll your own

Storage architecture diagrams

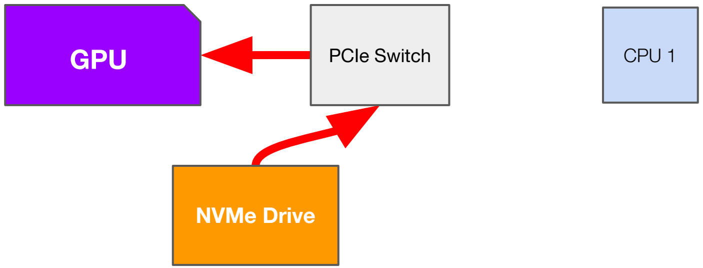


Single Tier NFS + local cache

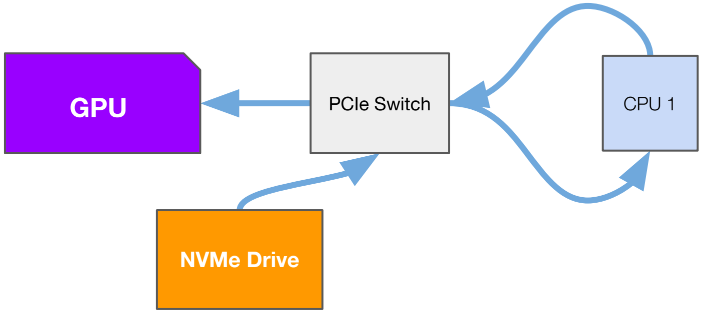
Parallel cluster file system + local cache

Tiered storage system + local cache

GPUDirect Storage



Data pathway with GPUDirect Storage
(Directly to the GPU via PCIe switch)



Data pathway without GPUDirect Storage
(Additional copy to CPU memory)

Storage choices

Open source storage options

FreeNAS™

OpenZFS

GlusterFS

lustre™

BeeGFS®

Proprietary storage options

NetApp™

PURESTORAGE®

ddn

WEKA

panasas®

SOFTIRON

Western Digital. ACTIVE SCALE

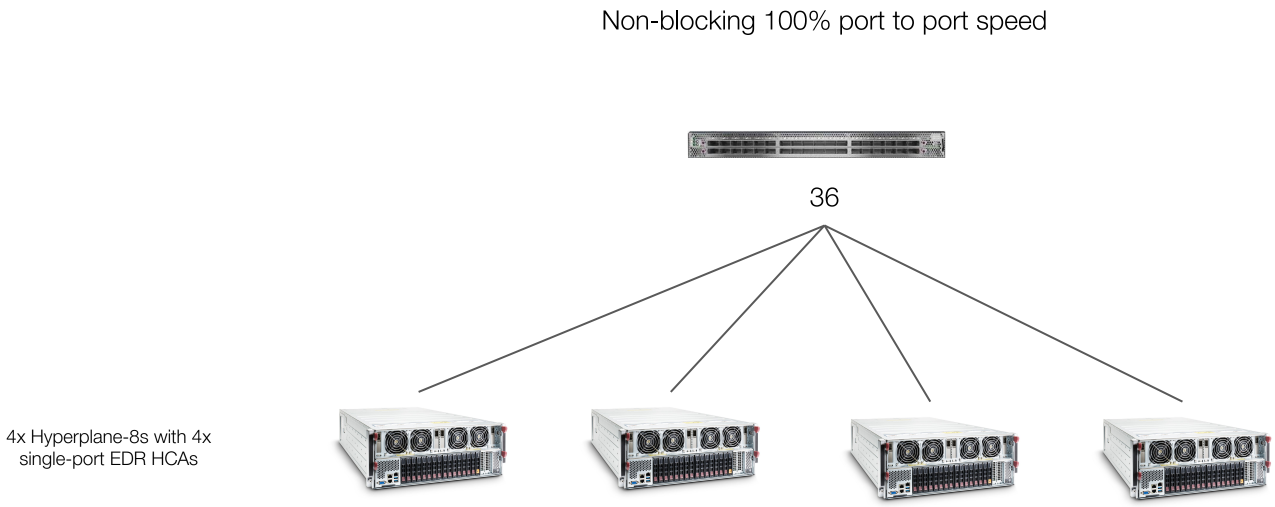
VAST

Cluster networking

Networking basics: switches

MSB7800 - 36 port EDR (100Gb/s) IB switch - 7.2Tb/s aggregate switch throughput
7.2 Tb/s / 36 ports = 7200 Gb/s / 36 ports = **200 Gb/s / port => (100 Gb/s EDR IB)**

Networking basics: single switch topology

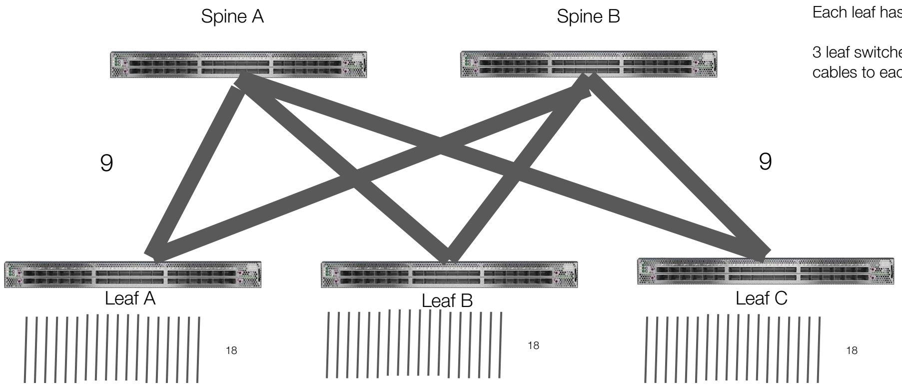


4*4 = 16 total IB cables

Networking basics: spine and leaf topology

To guarantee non blocking, ensure equal number of “southbound” and “northbound” ports on the leaf switches

Non-blocking fat tree topology



36/2 = 18 southbound and 18 northbound

Each leaf has 9 cables going northbound, 3 leaf switches

3 leaf switches * 9 northbound cables = 27 northbound leaf switch cables to each spine switch.

48 needed cables / 18 southbound = 2.666, round up to 3 => 3 leaf switches to provide full 48 ports.

54 total ports - we’re using 48/54.

12x Hyperplane-8s with 4x single-port EDR HCAs

4*12 = 48 total IB cables

Total southbound bandwidth of 5.4Tb/s available.

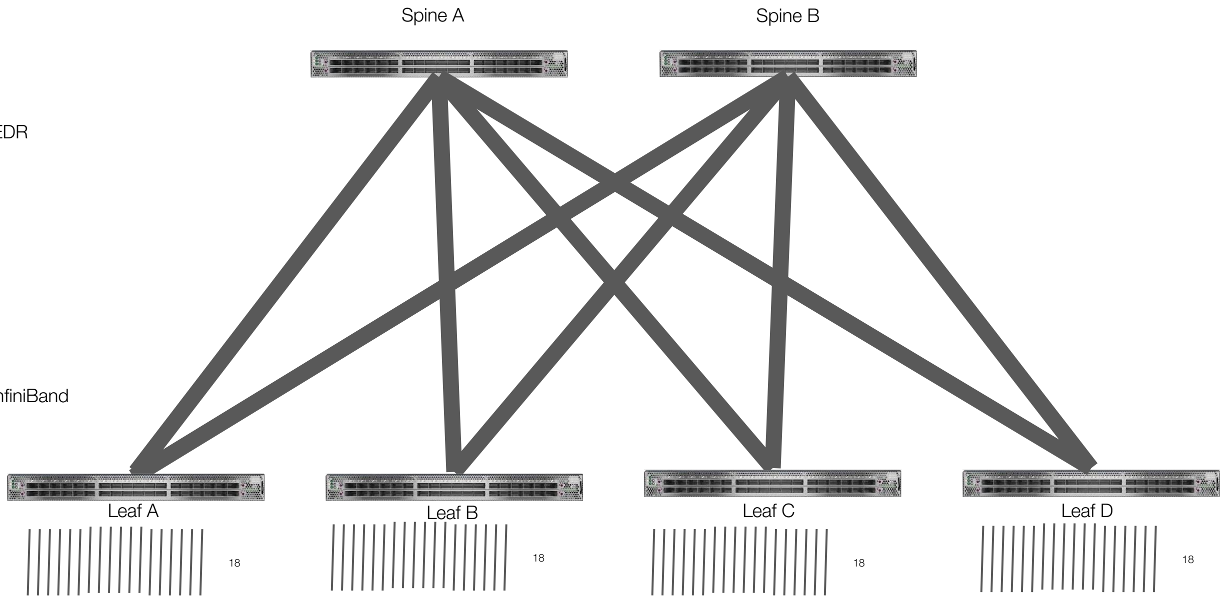
100% port-to-port bandwidth spine & leaf topology

Legend

A 36-port MSB7800 100Gb/s IB EDR Switch

Each thick grey line represents 9 InfiniBand cables.

Each thin black line represents 1 InfiniBand cable.

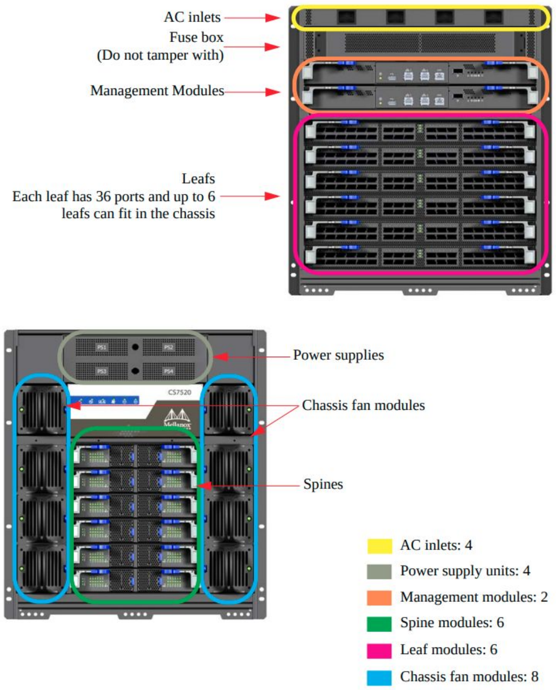


Networking basics: director switches

- Use a backplane instead of cables to build out the internal (spine <> leaf) network topology.
- To get 100% non-blocking bandwidth in the CS7520 you NEED to have all 6 spines installed.
- CS7520 has 216 southbound ports (and thus 216 northbound ports).
- Assuming each spine has only 36 ports, how many spines do you need to support 216 ports?
  - 36 ports / spine, 216 ports needed
  - 216 ports / 36 (ports / spine) = 6 spine switches

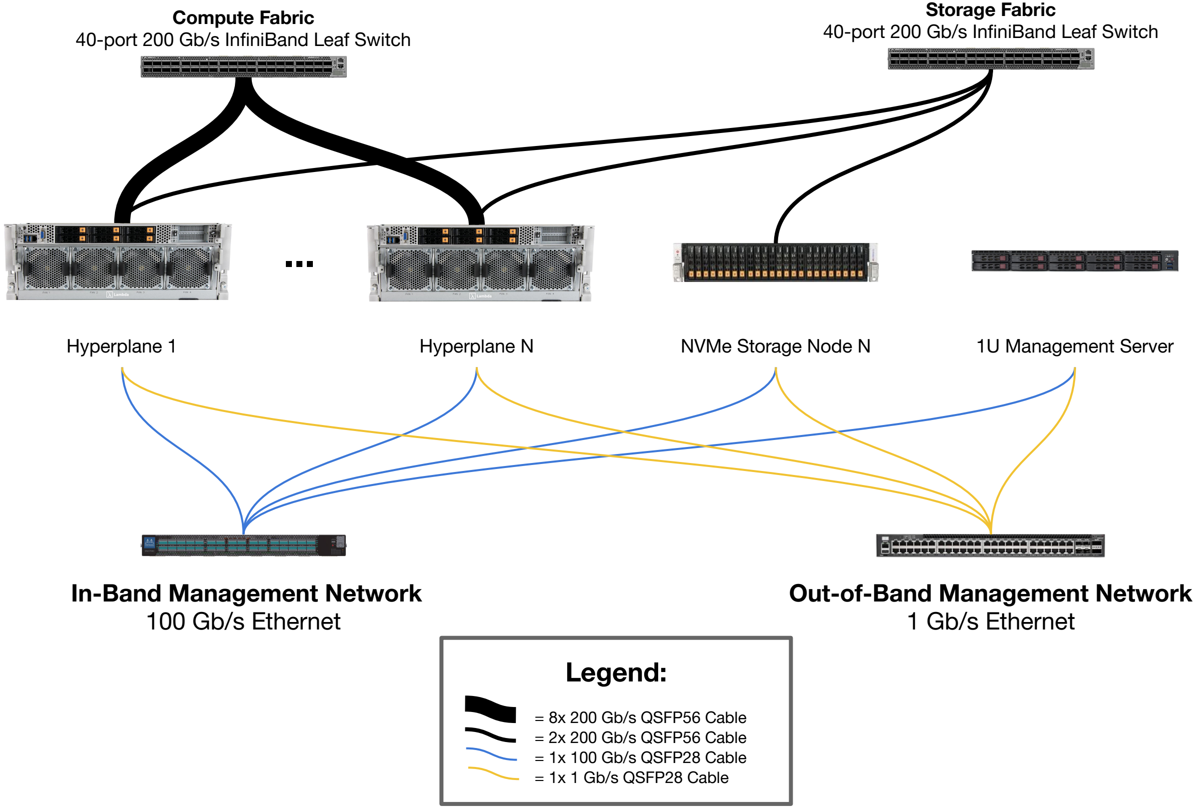
For more info see:
https://www.mellanox.com/related-docs/prod_ib_switch_systems/CS7520_Dismantling_Guide.pdf

Figure 1: Front and Rear View of the CS7520



Lambda Echelon Network Topology

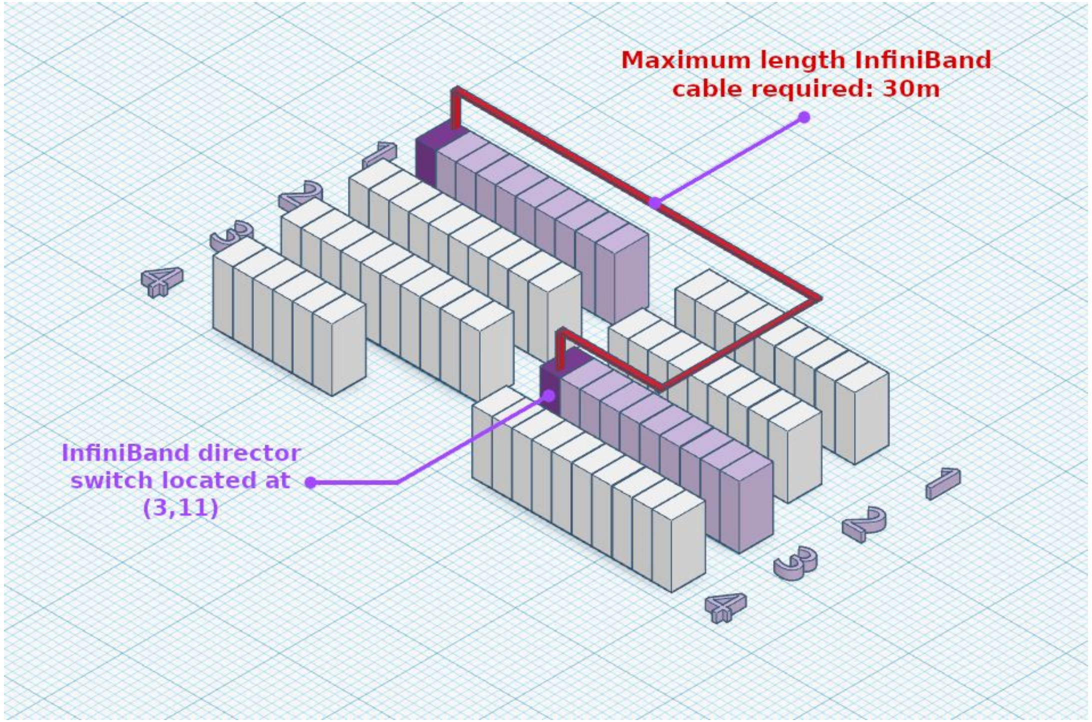
(Single Rack Configuration)



Data center floor planning & power distribution

Checklist:

- ❏ What rack does your DC provide?
- ❏ Max power (kW) per rack?
- ❏ What PDUs are provided, if any?
- ❏ What's the floor plan of the DC?
- ❏ Pull floor plan into DCIM/CAD to calculate cable lengths.



Software

There's a lot of software you'll want to run on your cluster. Some of this software is installed on the compute nodes, some of it is installed on the management nodes. Depending on your use case, these are some common packages you'll encounter.

- **Experiment management software** (Weights & Biases, Determined.ai, Spell.ml, etc.) - manage experiments and version control
- **Notebooks -** Jupyter Lab or iPython notebooks
- **Containers** - Docker mostly but also Singularity
- **Kubernetes** - manage instances and execute containers on them
- **Job scheduler** - (SLURM, Kubeflow)
- **Logging** - (Prometheus, Grafana, etc.)
- **Production inference** - (TensorFlow Serving, TensorRT)
- **Lambda Stack** - drivers, CUDA, TensorFlow, upgrade management, Dockerfile templates, etc.

Lambda Stack

GPU Accelerated Docker Containers:
Open Source Dockerfiles & NGC Container Support

System Wide Installation:
Works directly in your terminal or python notebook

Frameworks:

TensorFlow

PYTORCH

Caffe

Keras

Caffe2

Drivers:
NVIDIA Drivers, CUDA,
InfiniBand Drivers

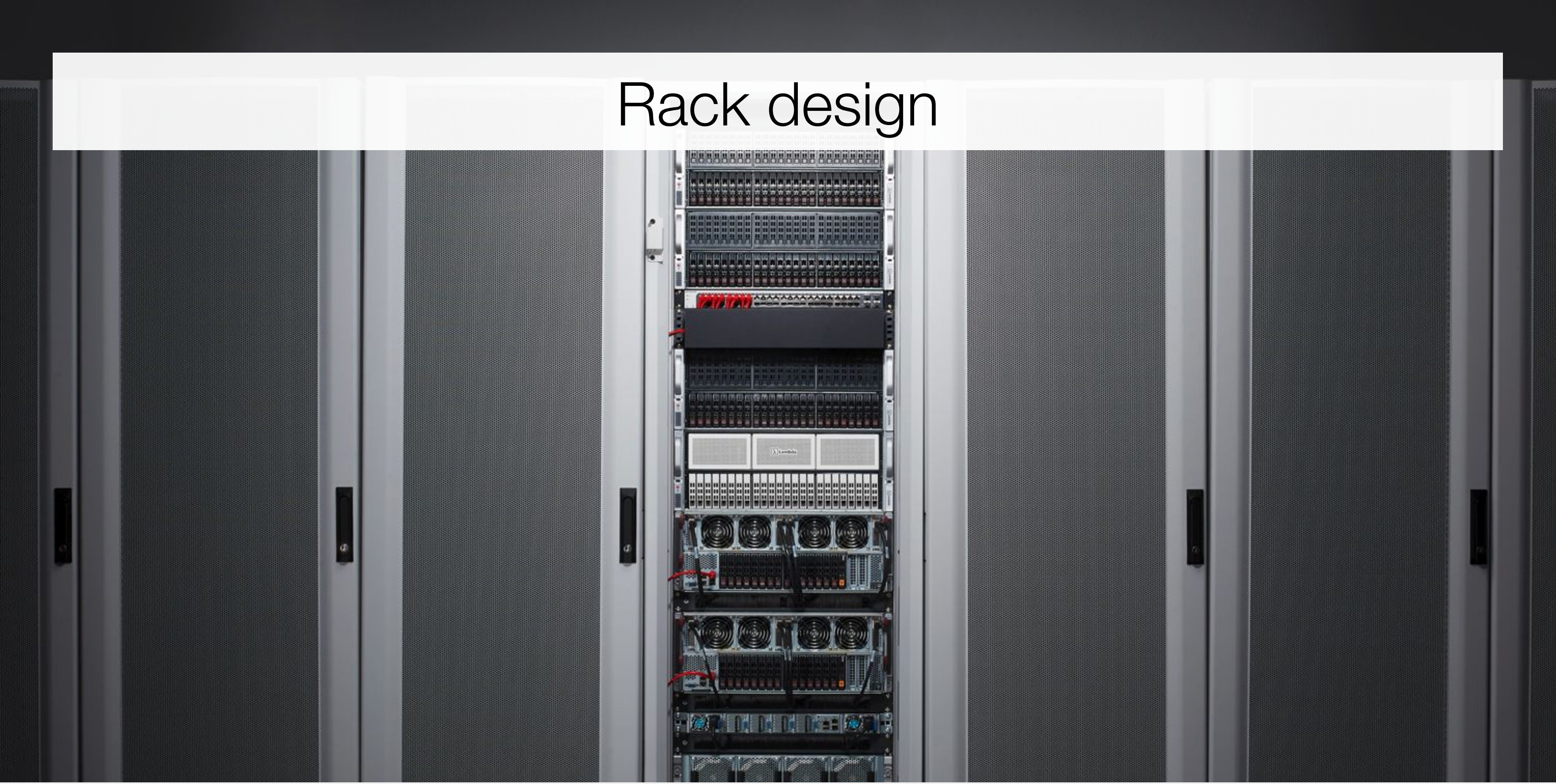
System Libraries:
protobuf, cuDNN, NCCL

Developer Tools:
git, vim, curl, emacs, tmux,
valgrind, ipython

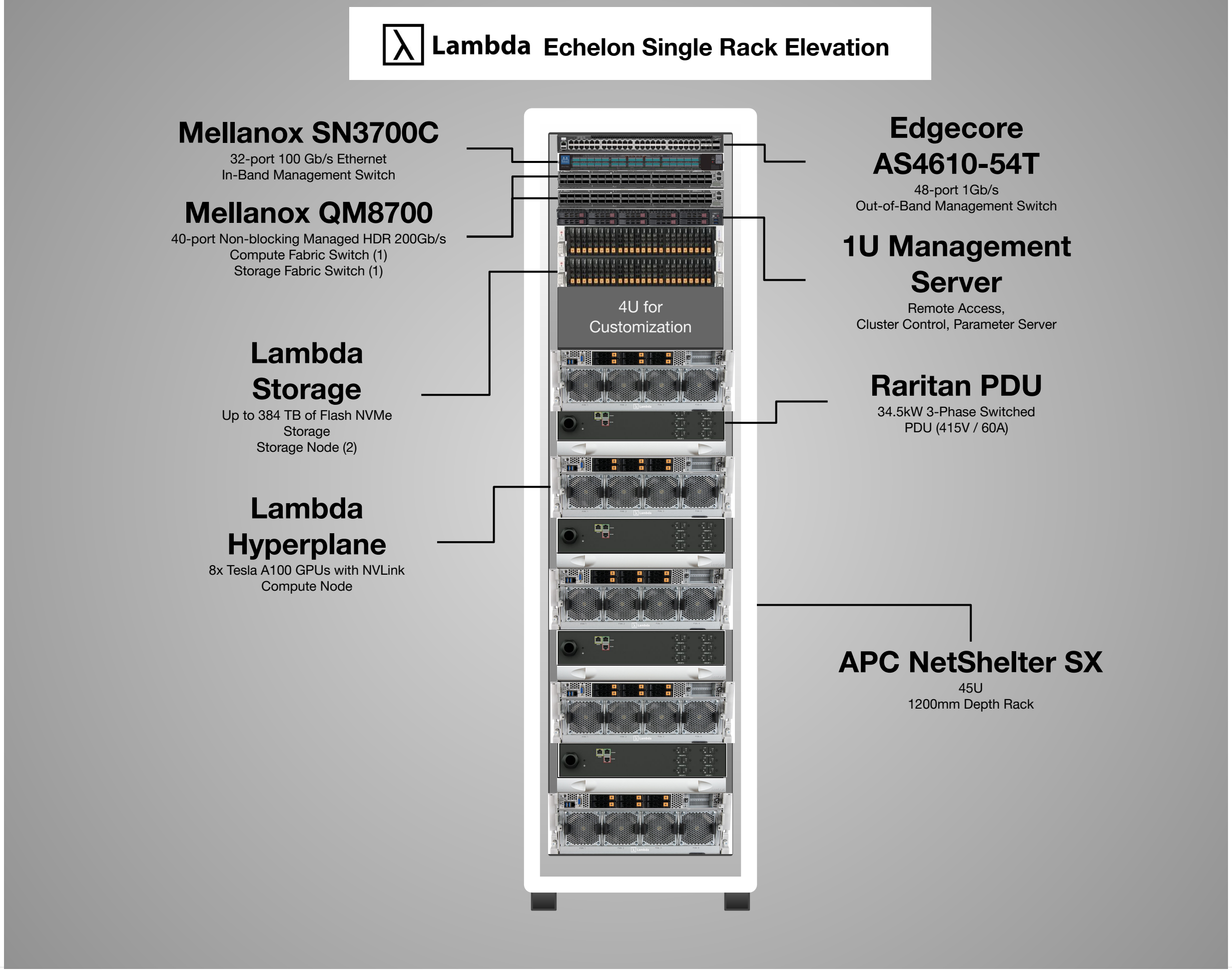
Operating System:
Ubuntu 20.04 LTS, Ubuntu 18.04 LTS or Ubuntu 16.04 LTS
(Long Term Support)

For more info on Lambda Stack, see: https://lambdalabs.com/lambda-stack-deep-learning-software

Rack design



Lambda


Lambda Echelon Single Rack Elevation
Mellanox SN3700C
32-port 100 Gb/s Ethernet
In-Band Management Switch
Mellanox QM8700
40-port Non-blocking Managed HDR 200Gb/s
Compute Fabric Switch (1)
Storage Fabric Switch (1)
Lambda Storage
Up to 384 TB of Flash NVMe Storage
Storage Node (2)
Lambda Hyperplane
8x Tesla A100 GPUs with NVLink
Compute Node
Edgecore AS4610-54T
48-port 1Gb/s
Out-of-Band Management Switch
1U Management Server
Remote Access,
Cluster Control, Parameter Server
Raritan PDU
34.5kW 3-Phase Switched
PDU (415V / 60A)
4U for Customization
APC NetShelter SX
45U
1200mm Depth Rack

Single Rack Cluster

	Compute, Storage, Networking Rack
1	**Edgecore** AS4610-54T - 1G IPMI Switch
2	**Mellanox** SN3700C - In Band Switch
3	**Mellanox** QM8700 - Compute Switch MGR
4	**Mellanox** QM8700 - Storage Switch MGR
5	**Lambda** - Management Node
6	**Lambda Matrix** - Flash Storage
7	Up to 384TB NVMe
8	**Lambda Matrix** - Flash Storage
9	Up to 384TB NVMe
10	**4U** Customizable Space
11	
12	
13	
14	**Lambda Hyperplane** 8x A100 Server
15	
16	
17	
18	**Raritan**
19	PX3-5008I2R-Q1 34.5kW PDU
20	1U Blanking Panel
21	**Lambda Hyperplane** 8x A100 Server
22	
23	
24	
25	**Raritan**
26	PX3-5008I2R-Q1 34.5kW PDU
27	1U Blanking Panel
28	**Lambda Hyperplane** 8x A100 Server
29	
30	
31	
32	**Raritan**
33	PX3-5008I2R-Q1 34.5kW PDU
34	1U Blanking Panel
35	**Lambda Hyperplane** 8x A100 Server
36	
37	
38	
39	**Raritan**
40	PX3-5008I2R-Q1 34.5kW PDU
41	1U Blanking Panel
42	**Lambda Hyperplane** 8x A100 Server
43	
44	
45	

Lambda Echelon - Single Rack Cluster - Bill of Materials

Description	Qty	TDP Draw (kW)	Ext. TDP (kW)
Lambda Hyperplane - 8x A100 Server	5	6.5	32.5
1U Management Node	1	1	1
2U Flash Storage Server	2	0.75	1.5
Mellanox SN3700C - 32 port - 100 Gb/s Ethernet - In-Band Switch	1	0.25	0.25
Mellanox QM8700 - 40 port - 200 Gb/s IB- Compute Fabric Switch	1	0.8	0.8
Mellanox QM8700 - 40 port - 200 Gb/s IB - Storage Fabric Switch	1	0.8	0.8
Edgecore AS4610-54T - 48 port - 1 Gb/s Ethernet - IPMI Switch	1	0.1	0.1
Mellanox MFS1S00-H003E - 3m active 200 Gb/s IB cables	49	0	0
Mellanox MFA1A00-C003 - 3m active 100 Gb/s Ethernet cables	8	0	0
Cat 5e Ethernet Cables - 1m/2m/3m	16	0	0
Line A - Raritan PX3-5008I2R-Q1 - 34.5kW PDU 415V 60A Input IEC 60309 3P+N+E 6h 60A (4P5W)	2	0	0
(Optional Line B) Raritan PX3-5008I2R-Q1 - 34.5kW PDU 415V 60A Input IEC 60309 3P+N+E 6h 60A (4P5W)	2	0	0
1U Blanking Panel	4	0	0
APC Netshelter SX 3305W - 45U - 600mm x 1200mm Rack	1	0	0
C14 to C13 cables (2 ft)	14	0	0
C13 to C20 cables (2 ft)	20	0	0
		Rack TDP Draw (kW):	36.95

Lambda Echelon - Single Rack Cluster - Bill of Materials

The most important number on this page: **Total rack TDP**.

Description	Qty	TDP Draw (kW)	Ext. TDP (kW)
Lambda Hyperplane - 8x A100 Server	5	6.5	32.5
1U Management Node	1	1	1
2U Flash Storage Server	2	0.75	1.5
Mellanox SN3700C - 32 port - 100 Gb/s Ethernet - In-Band Switch	1	0.25	0.25
Mellanox QM8700 - 40 port - 200 Gb/s IB- Compute Fabric Switch	1	0.8	0.8
Mellanox QM8700 - 40 port - 200 Gb/s IB - Storage Fabric Switch	1	0.8	0.8
Edgecore AS4610-54T - 48 port - 1 Gb/s Ethernet - IPMI Switch	1	0.1	0.1
Mellanox MFS1S00-H003E - 3m active 200 Gb/s IB cables	49	0	0
Mellanox MFA1A00-C003 - 3m active 100 Gb/s Ethernet cables	8	0	0
Cat 5e Ethernet Cables - 1m/2m/3m	16	0	0
Line A - Raritan PX3-5008I2R-Q1 - 34.5kW PDU 415V 60A Input IEC 60309 3P+N+E 6h 60A (4P5W)	2	0	0
(Optional Line B) Raritan PX3-5008I2R-Q1 - 34.5kW PDU 415V 60A Input IEC 60309 3P+N+E 6h 60A (4P5W)	2	0	0
1U Blanking Panel	1	0	0
APC Netshelter SX 3305W - 45U - 600mm x 1200mm Rack	1	0	0
C14 to C13 cables (2 ft)	14	0	0
C13 to C20 cables (2 ft)	20	0	0
		Rack TDP Draw (kW):	36.95

Back to basics

Single phase systems

P (watts) = V (volts) * I (amps)

We use *I* from the French, *intensité du courant*.
Because that's what André-Marie Ampère used.

3-phase systems

P = 3 * V / sqrt(3) * I

This simplifies to P = sqrt(3) * V * I
Because 3 / sqrt(3) = sqrt(3)

Real life 3-phase systems

P = sqrt(3) * V * I * 0.8

It's very common to see an 80% regulatory derating factor applied to PDUs.

How do PDU manufacturers calculate power capacity?

Example PDU:
APC 8966

From the APC8966 Data Sheet:

Input frequency
50/60 Hz

Number of Power Cords
1

Load Capacity
17300VA

Maximum Input Current
60A

Maximum Line Current
48A

Regulatory Derated Input Current (North America)
48A

Nominal Output Voltage
208V

Nominal Input Voltage
208V 3PH

Input Connections
IEC 60309 60A 3P + PE

P = sqrt(3) * V * I * 0.8

Plug in the numbers from the data sheet:

P = sqrt(3) * **208** * **60** * **0.8**
P = 17292.7953

<u>P = 17.3kVA</u>

*See how they derate the maximum input current of 60A to the "Regulatory Derated Input Current (North America)" 48A? That's 48A = 60A * 0.8.*

That's where the 0.8 came from in the previous slide.

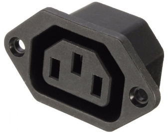
Some common PDU input plugs

IEC60309 - 60A 3-phase plug - 208V
Blue means the system is between 200 and 250V.

IEC60309 - 60A 3-phase plug - 415V
Red means the system is above 400V.

NEMA L15-30P 30A 3-phase plug - 208V

Receptacles & Plugs

Plug Photo	Name	Plugs into	Receptacle Photo	Max Amps
	IEC C13 Plug	IEC C14 Receptacle (on server)		**15A** (Max power ~2.5kW)
	IEC C14 Plug	IEC C13 Receptacle (on PDU)		**15A** (Max power ~2.5kW)
	IEC C19 Plug	IEC C20 Receptacle (on server)		**20A** (Max power ~3.2kW)
	IEC C20 Plug	IEC C19 Receptacle (on PDU)		**20A** (Max power ~3.2kW)

***IEC** stands for International Electrotechnical Commission, an international standards organization headquartered in Switzerland.*

Lambda Echelon - Single Rack Cluster - Bill of Materials

Description	Qty	TDP Draw (kW)	Ext. TDP (kW)
Lambda Hyperplane - 8x A100 Server	5	6.5	32.5
1U Management Node	1	1	1
2U Flash Storage Server	2	0.75	1.5
Mellanox SN3700C - 32 port - 100 Gb/s Ethernet - In-Band Switch	1	0.25	0.25
Mellanox QM8700 - 40 port - 200 Gb/s IB- Compute Fabric Switch	1	0.8	0.8
Mellanox QM8700 - 40 port - 200 Gb/s IB - Storage Fabric Switch	1	0.8	0.8
Edgecore AS4610-54T - 48 port - 1 Gb/s Ethernet - IPMI Switch	1	0.1	0.1
Mellanox MFS1S00-H003E - 3m active 200 Gb/s IB cables	49	0	0
Mellanox MFA1A00-C003 - 3m active 100 Gb/s Ethernet cables	8	0	0
Cat 5e Ethernet Cables - 1m/2m/3m	16	0	0
Line A - Raritan PX3-5008I2R-Q1 - 34.5kW PDU 415V 60A Input IEC 60309 3P+N+E 6h 60A (4P5W)	2	0	0
(Optional Line B) Raritan PX3-5008I2R-Q1 - 34.5kW PDU 415V 60A Input IEC 60309 3P+N+E 6h 60A (4P5W)	2	0	0
1U Blanking Panel	1	0	0
APC Netshelter SX 3305W - 45U - 600mm x 1200mm Rack	1	0	0
C14 to C13 cables (2 ft)	14	0	0
C13 to C20 cables (2 ft)	20	0	0
		Rack TDP Draw (kW):	36.95

Now you understand how much work goes into this bill of materials.

Lambda Echelon rack elevations

	Single Rack Cluster
	Compute, Storage, Networking Rack
1	**Edgecore** AS4610-54T - 1G IPMI Switch
2	**Mellanox** SN3700C - In Band Switch
3	**Mellanox** QM8700 - Compute Switch MGR
4	**Mellanox** QM8700 - Storage Switch MGR
5	**Lambda** - Management Node
6	**Lambda Matrix** - Flash Storage
7	Up to 384TB NVMe
8	**Lambda Matrix** - Flash Storage
9	Up to 384TB NVMe
10	**4U**
11	Customizable
12	Space
13	
14	
15	**Lambda Hyperplane**
16	8x A100 Server
17	
18	**Raritan**
19	PX3-5008I2R-Q1 34.5kW PDU
20	1U Blanking Panel
21	
22	**Lambda Hyperplane**
23	8x A100 Server
24	
25	**Raritan**
26	PX3-5008I2R-Q1 34.5kW PDU
27	1U Blanking Panel
28	
29	**Lambda Hyperplane**
30	8x A100 Server
31	
32	**Raritan**
33	PX3-5008I2R-Q1 34.5kW PDU
34	1U Blanking Panel
35	
36	**Lambda Hyperplane**
37	8x A100 Server
38	
39	**Raritan**
40	PX3-5008I2R-Q1 34.5kW PDU
41	1U Blanking Panel
42	
43	**Lambda Hyperplane**
44	8x A100 Server
45	

40 GPUs

Training time for Mask R-CNN on MSCOCO: 25 mins

Source: https://mlperf.org/training-results-0-7/

Four Rack Cluster

U	(1) Compute + Network + Storage Rack	(1) Compute + Networking Rack	(1) Compute + Storage Rack	(1) Compute + Storage Rack
1	Edgecore AS4610-54T - 1G IPMI Switch	**Edgecore** AS4610-54T - 1G IPMI Switch	**Edgecore** AS4610-54T - 1G IPMI Switch	**Edgecore** AS4610-54T - 1G IPMI Switch
2	**Mellanox** SN3700C - In-Band Switch	**Mellanox** QM8700 - Spine Switch MGR	**Lambda Matrix** - Flash Storage	**Lambda Matrix** - Flash Storage
3	**Mellanox** QM8700 - Storage Switch	**Mellanox** QM8790 - Spine Switch	Up to 384TB NVMe	Up to 384TB NVMe
4	**Lambda** - Management Node	**Mellanox** QM8790 - Spine Switch	**Lambda Matrix** - Flash Storage	**Lambda Matrix** - Flash Storage
5	**Lambda Matrix** - Flash Storage	**Mellanox** QM8790 - Spine Switch	Up to 384TB NVMe	Up to 384TB NVMe
6	Up to 384TB NVMe	**Mellanox** QM8790 - Leaf Switch		
7	**Lambda Matrix** - Flash Storage	**Mellanox** QM8790 - Leaf Switch		
8	Up to 384TB NVMe	**Mellanox** QM8790 - Leaf Switch		
9		**Mellanox** QM8790 - Leaf Switch	**8U**	**8U**
10	**5U**	**Mellanox** QM8790 - Leaf Switch	Customizable	Customizable
11	Customizable	**Mellanox** QM8790 - Leaf Switch	Space	Space
12	Space	**Mellanox** QM8790 - Leaf Switch		
13		**Mellanox** QM8790 - Leaf Switch		
14				
15	**Lambda Hyperplane**	**Lambda Hyperplane**	**Lambda Hyperplane**	**Lambda Hyperplane**
16	8x A100 Server	8x A100 Server	8x A100 Server	8x A100 Server
17				
18	**Raritan**	**Raritan**	**Raritan**	**Raritan**
19	PX3-5008I2R-Q1 34.5kW PDU	PX3-5008I2R-Q1 34.5kW PDU	PX3-5008I2R-Q1 34.5kW PDU	PX3-5008I2R-Q1 34.5kW PDU
20	1U Blanking Panel	1U Blanking Panel	1U Blanking Panel	1U Blanking Panel
21				
22	**Lambda Hyperplane**	**Lambda Hyperplane**	**Lambda Hyperplane**	**Lambda Hyperplane**
23	8x A100 Server	8x A100 Server	8x A100 Server	8x A100 Server
24				
25	**Raritan**	**Raritan**	**Raritan**	**Raritan**
26	PX3-5008I2R-Q1 34.5kW PDU	PX3-5008I2R-Q1 34.5kW PDU	PX3-5008I2R-Q1 34.5kW PDU	PX3-5008I2R-Q1 34.5kW PDU
27	1U Blanking Panel	1U Blanking Panel	1U Blanking Panel	1U Blanking Panel
28				
29	**Lambda Hyperplane**	**Lambda Hyperplane**	**Lambda Hyperplane**	**Lambda Hyperplane**
30	8x A100 Server	8x A100 Server	8x A100 Server	8x A100 Server
31				
32	**Raritan**	**Raritan**	**Raritan**	**Raritan**
33	PX3-5008I2R-Q1 34.5kW PDU	PX3-5008I2R-Q1 34.5kW PDU	PX3-5008I2R-Q1 34.5kW PDU	PX3-5008I2R-Q1 34.5kW PDU
34	1U Blanking Panel	1U Blanking Panel	1U Blanking Panel	1U Blanking Panel
35				
36	**Lambda Hyperplane**	**Lambda Hyperplane**	**Lambda Hyperplane**	**Lambda Hyperplane**
37	8x A100 Server	8x A100 Server	8x A100 Server	8x A100 Server
38				
39	**Raritan**	**Raritan**	**Raritan**	**Raritan**
40	PX3-5008I2R-Q1 34.5kW PDU	PX3-5008I2R-Q1 34.5kW PDU	PX3-5008I2R-Q1 34.5kW PDU	PX3-5008I2R-Q1 34.5kW PDU
41	1U Blanking Panel	1U Blanking Panel	1U Blanking Panel	1U Blanking Panel
42				
43	**Lambda Hyperplane**	**Lambda Hyperplane**	**Lambda Hyperplane**	**Lambda Hyperplane**
44	8x A100 Server	8x A100 Server	8x A100 Server	8x A100 Server
45				

160 GPUs

Training time for Transformer WMT E-G Big: ~1.02 mins

Source:
https://mlperf.org/training-results-0-7/

21 Rack Cluster

	(15) Compute + Storage Rack		(1) Core Networking Rack		(5) Compute + Networking Rack
1	Edgecore AS4610-54T - 1G IPMI Switch	1	Edgecore AS4610-54T - 1G IPMI Switch	1	Edgecore AS4610-54T - 1G IPMI Switch
2	1U Blanking Panel	2	1U Blanking Panel	2	Mellanox QM8790 - Leaf Switch
3	1U Blanking Panel	3	Mellanox QM8700 - Spine Switch MGR	3	Mellanox QM8790 - Leaf Switch
4	Lambda Matrix - Flash Storage	4	Mellanox QM8790 - Spine Switch	4	Mellanox QM8790 - Leaf Switch
5	Up to 384TB NVMe	5	Mellanox QM8790 - Spine Switch	5	Mellanox QM8790 - Leaf Switch
6	Lambda Matrix - Flash Storage	6	Mellanox QM8790 - Spine Switch	6	Mellanox QM8790 - Leaf Switch
7	Up to 384TB NVMe	7	1U Blanking Panel	7	Mellanox QM8790 - Leaf Switch
8		8	Mellanox QM8790 - Spine Switch	8	Mellanox QM8790 - Leaf Switch
9	6U	9	Mellanox QM8790 - Spine Switch	9	Mellanox QM8790 - Leaf Switch
10	Customizable	10	Mellanox QM8790 - Spine Switch	10	Mellanox QM8790 - Storage Leaf Switch
11	Space	11	Mellanox QM8790 - Spine Switch	11	Mellanox QM8790 - Storage Leaf Switch
12		12	1U Blanking Panel	12	Mellanox SN3700C - In-Band Switch
13		13	Mellanox QM8790 - Spine Switch	13	1U Blanking Panel
14		14	Mellanox QM8790 - Spine Switch	14	
15	Lambda Hyperplane	15	Mellanox QM8790 - Spine Switch	15	Lambda Hyperplane
16	8x A100 Server	16	Mellanox QM8790 - Spine Switch	16	8x A100 Server
17		17	1U Blanking Panel	17	
18	Raritan	18	Mellanox QM8790 - Spine Switch	18	Raritan
19	PX3-5008I2R-Q1 34.5kW PDU	19	Mellanox QM8790 - Spine Switch	19	PX3-5008I2R-Q1 34.5kW PDU
20	1U Blanking Panel	20	Mellanox QM8790 - Spine Switch	20	1U Blanking Panel
21		21	Mellanox QM8790 - Spine Switch	21	
22	Lambda Hyperplane	22	1U Blanking Panel	22	Lambda Hyperplane
23	8x A100 Server	23	Mellanox QM8790 - Spine Switch	23	8x A100 Server
24		24	Mellanox QM8790 - Spine Switch	24	
25	Raritan	25	Mellanox QM8790 - Spine Switch	25	Raritan
26	PX3-5008I2R-Q1 34.5kW PDU	26	Mellanox QM8790 - Spine Switch	26	PX3-5008I2R-Q1 34.5kW PDU
27	1U Blanking Panel	27	1U Blanking Panel	27	1U Blanking Panel
28		28	Mellanox QM8700 - Storage Spine Switch MGR	28	
29	Lambda Hyperplane	29	Mellanox QM8790 - Storage Spine Switch	29	Lambda Hyperplane
30	8x A100 Server	30	Mellanox QM8790 - Storage Spine Switch	30	8x A100 Server
31		31	Mellanox QM8790 - Storage Spine Switch	31	
32	Raritan	32	Mellanox QM8790 - Storage Spine Switch	32	Raritan
33	PX3-5008I2R-Q1 34.5kW PDU	33	1U Blanking Panel	33	PX3-5008I2R-Q1 34.5kW PDU
34	1U Blanking Panel	34	Lambda - Management Node	34	1U Blanking Panel
35		35	Lambda - Management Node	35	
36	Lambda Hyperplane	36	Lambda - Management Node	36	Lambda Hyperplane
37	8x A100 Server	37	Lambda - Management Node	37	8x A100 Server
38		38	Lambda - Management Node	38	
39	Raritan	39	Mellanox SN3700C - In-Band Spine	39	Raritan
40	PX3-5008I2R-Q1 34.5kW PDU	40	1U Blanking Panel	40	PX3-5008I2R-Q1 34.5kW PDU
41	1U Blanking Panel	41	1U Blanking Panel	41	1U Blanking Panel
42		42	1U Blanking Panel	42	
43	Lambda Hyperplane	43	1U Blanking Panel	43	Lambda Hyperplane
44	8x A100 Server	44	1U Blanking Panel	44	8x A100 Server
45		45	1U Blanking Panel	45	

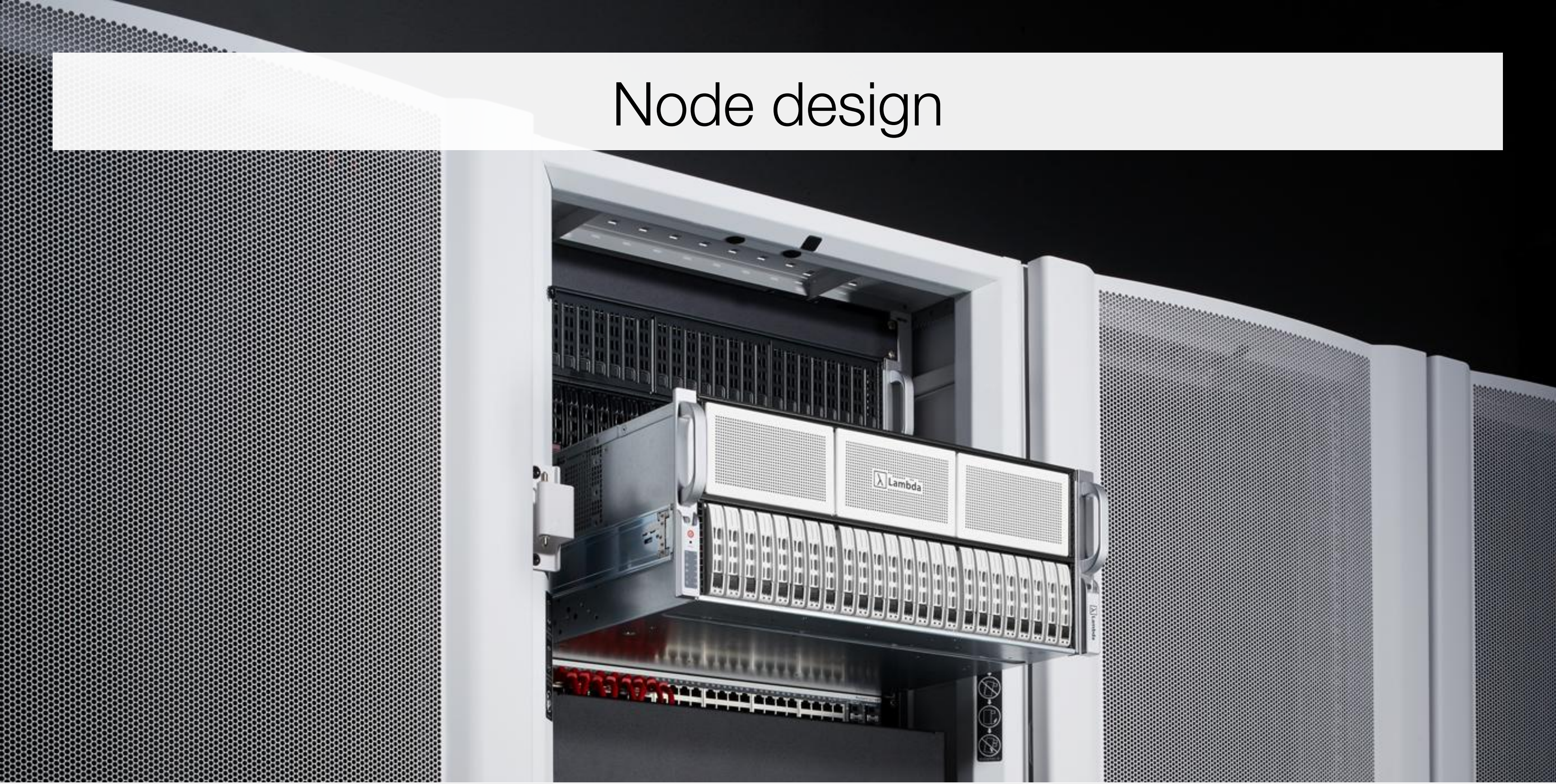
800 GPUs

Training time for ResNet 50 on ImageNet:
1.06 mins

Source:
https://mlperf.org/training-results-0-7/
https://github.com/NVIDIA/DeepLearningExamples/blob/2de29455a696b5608e24daa95f579f689b2d59e1/PyTorch/Translation/Transformer/README.md

Lambda

Node design



Choosing the right GPU depends on your use case

Quadro RTX 8000

NVIDIA A100

NVIDIA T4

Hyperparameter Search	Large scale distributed training	Production inference

GPU benchmarking is hard!

Lambda Benchmarks

https://lambdalabs.com/blog/

MLPerf

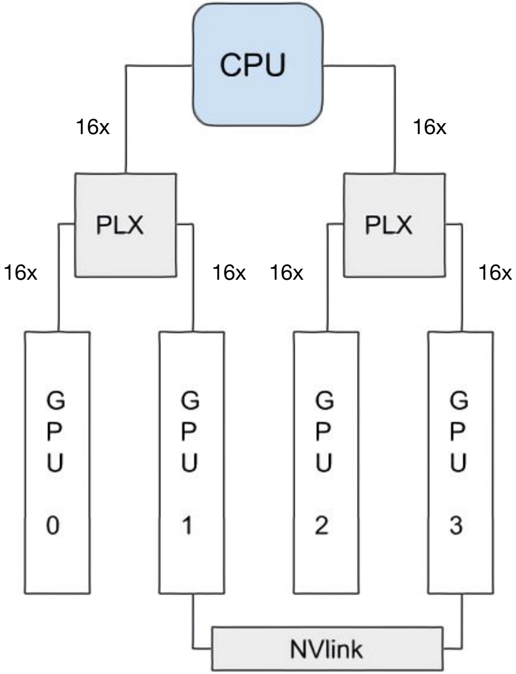
https://mlperf.org

What to look for in a node / compute system

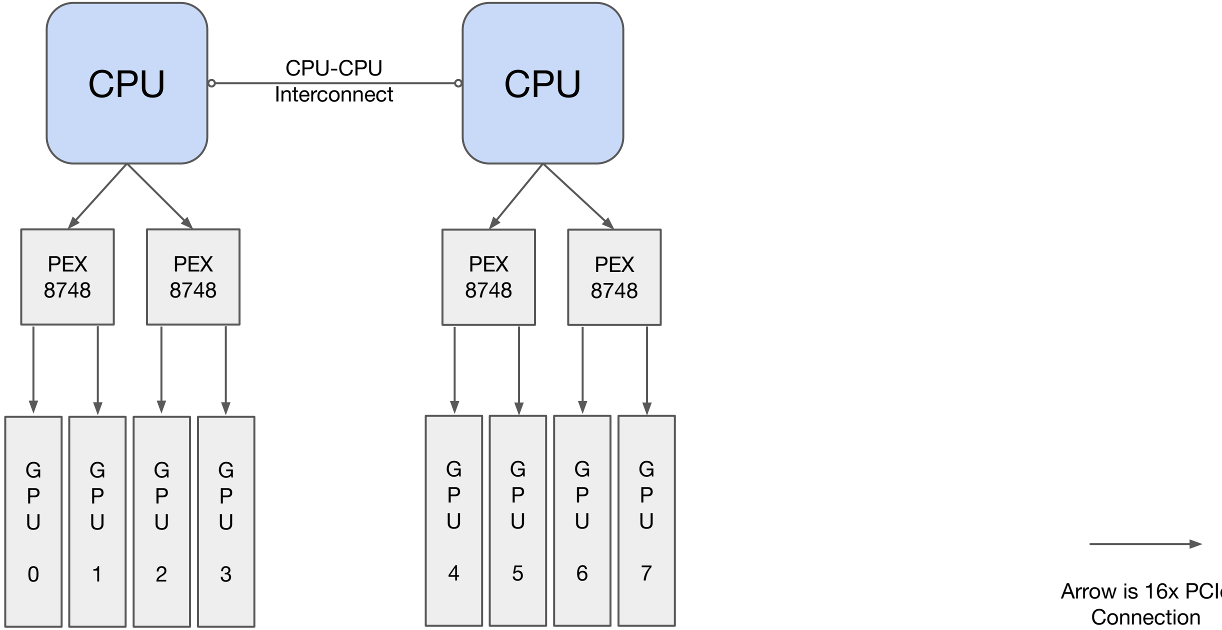
1. Number of PCIe lanes. (Affects total bandwidth.)
2. PCIe topology. (Tradeoff between CPU<>GPU and GPU<>GPU transfer.)
3. PCIe generation. (Gen 4 is 2x bandwidth of Gen 3.)
4. NUMA node topology. (Affects GPU peering & virtualization.)
5. FLOPS / $.

GPU peering & PCIe topology

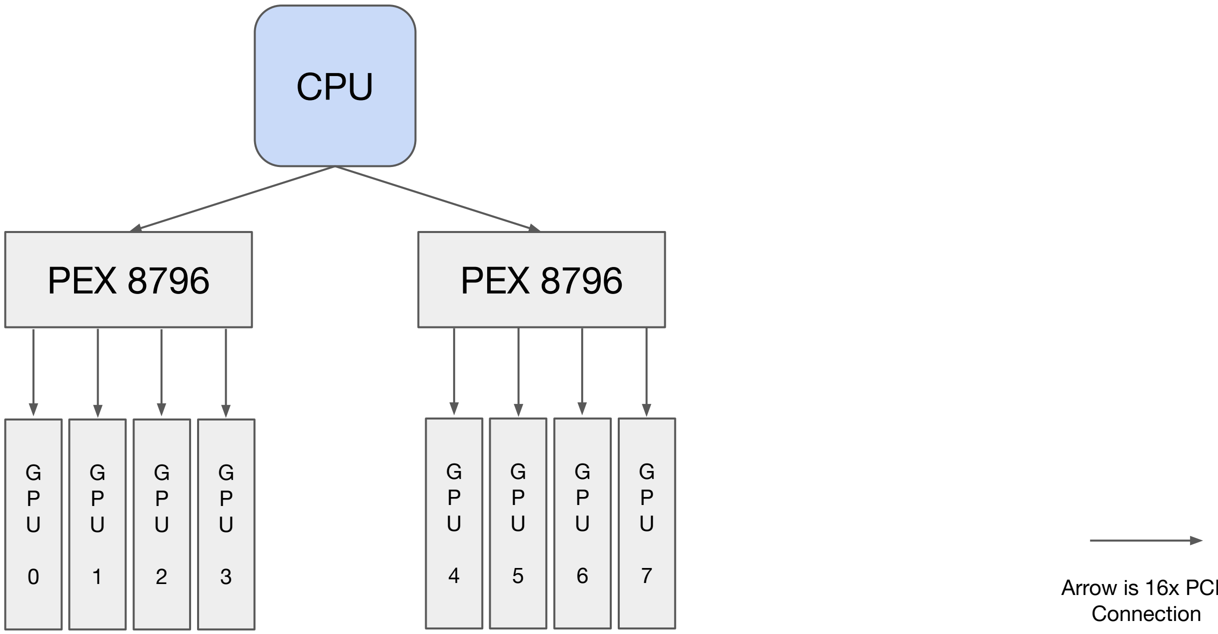
Example PCIe topology



Dual root PCIe topology

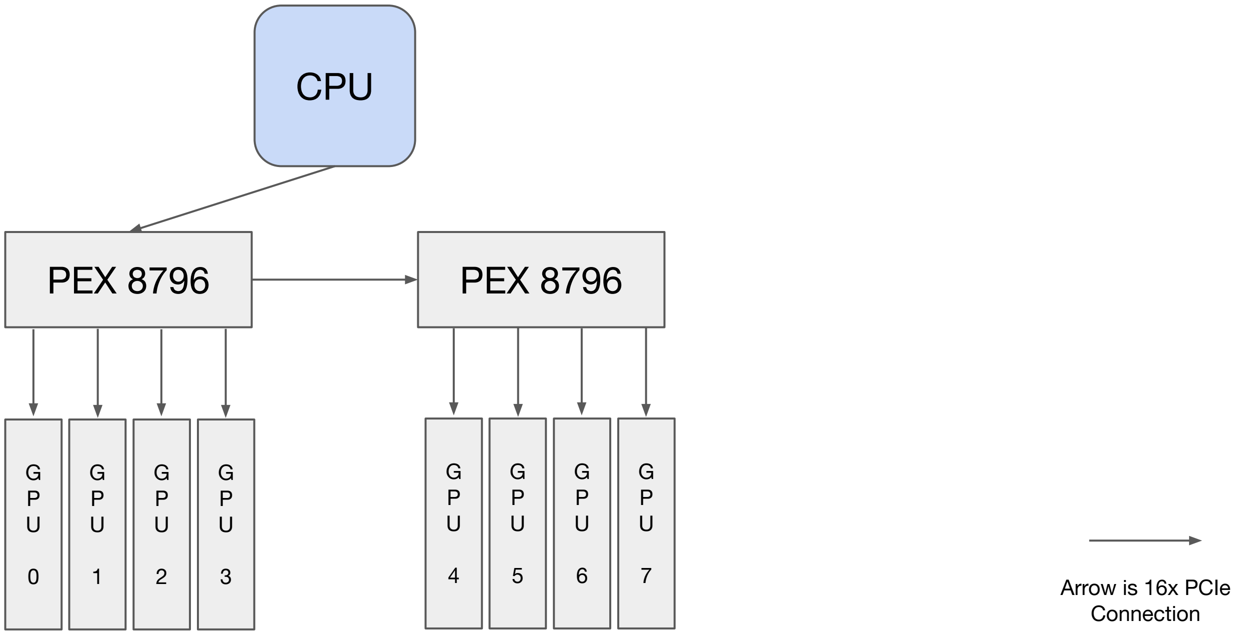


Single root PCIe topology

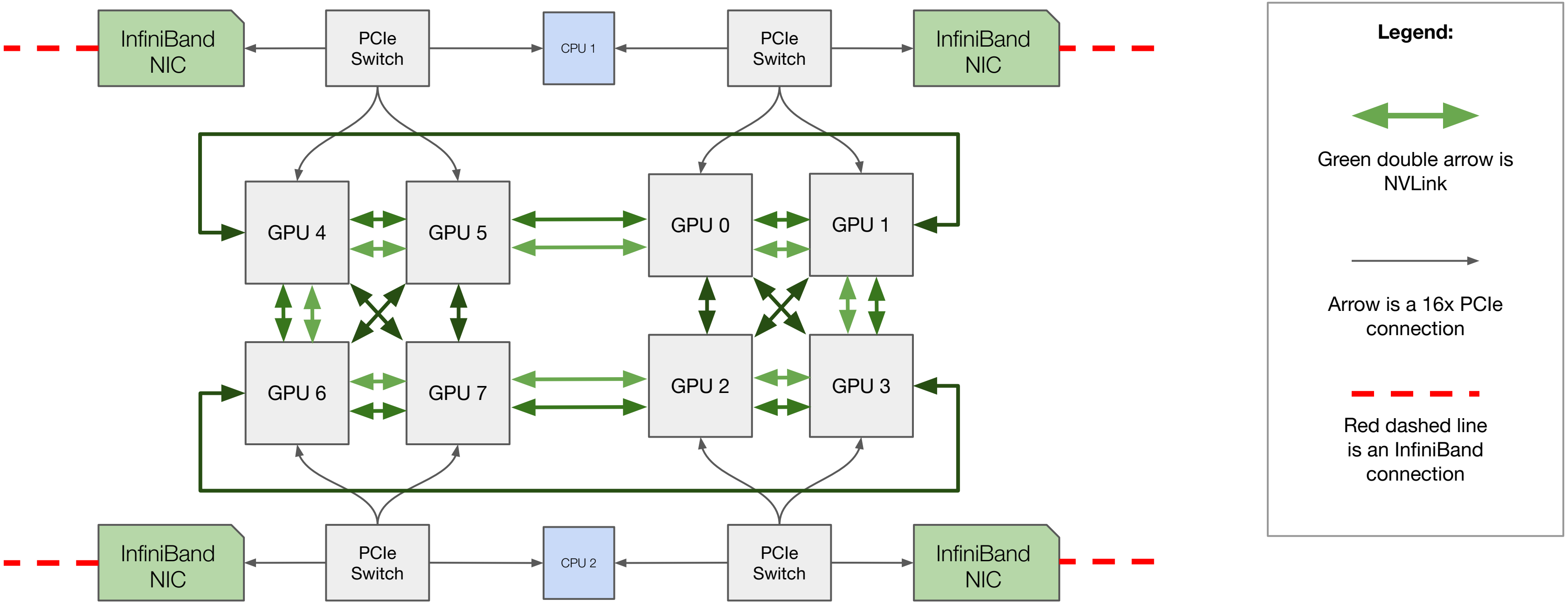


Source: Lambda

Cascaded PCIe topology

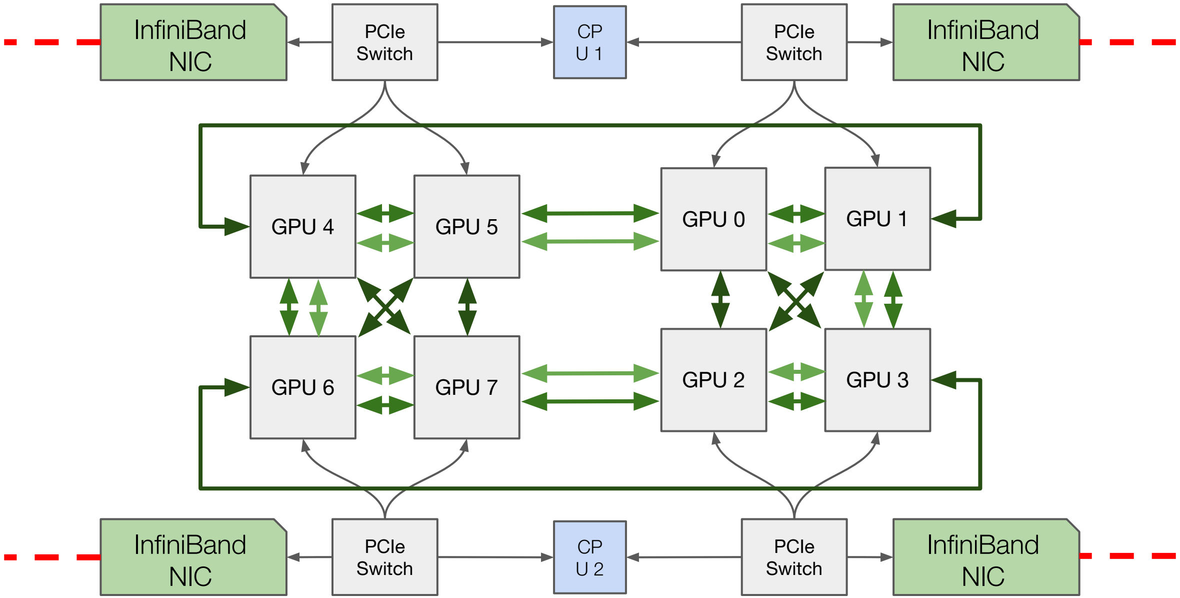


NVLink hybrid cube-mesh topology



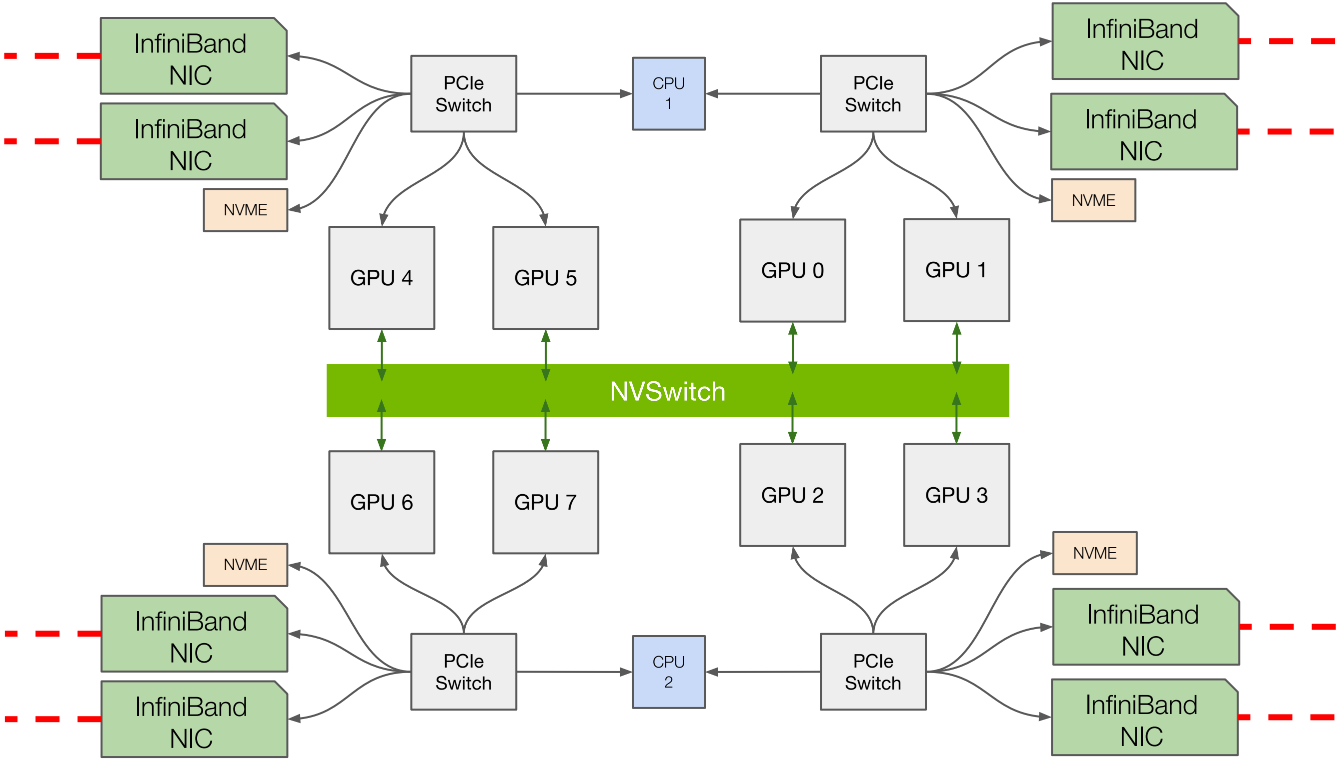
Source: Lambda

Four non-overlapping pathways

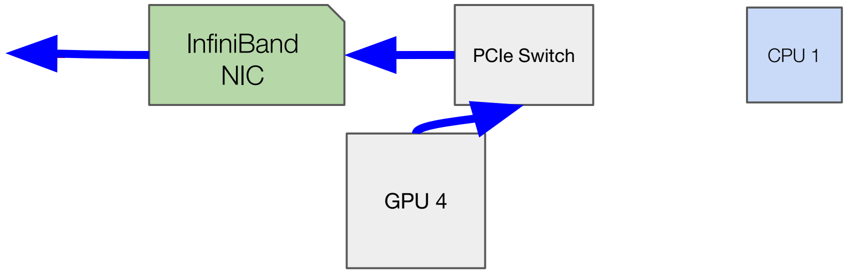


Each GPU has six NVLink connections which allows for four non-overlapping pathways to be drawn through all eight GPUs on the system and out an InfiniBand card, optimizing both GPU to GPU and node to node communication during distributed all-reduce. (See Sylvain Jeaugey's "NCCL 2.0" presentation for more information.)

NVSwitch Topology

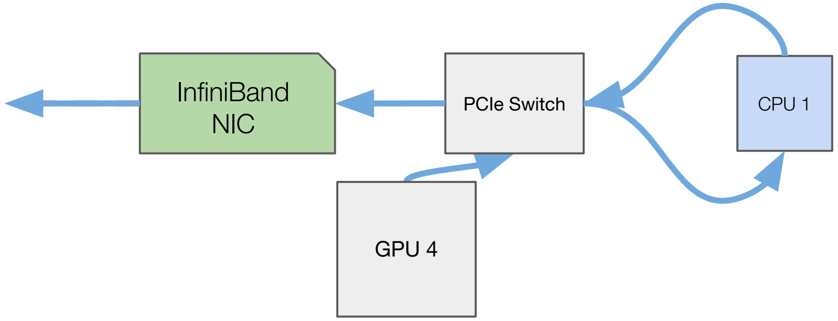


GPUDirect RDMA over InfiniBand



Data pathway with RDMA
(Directly out the door via PCIe switch)

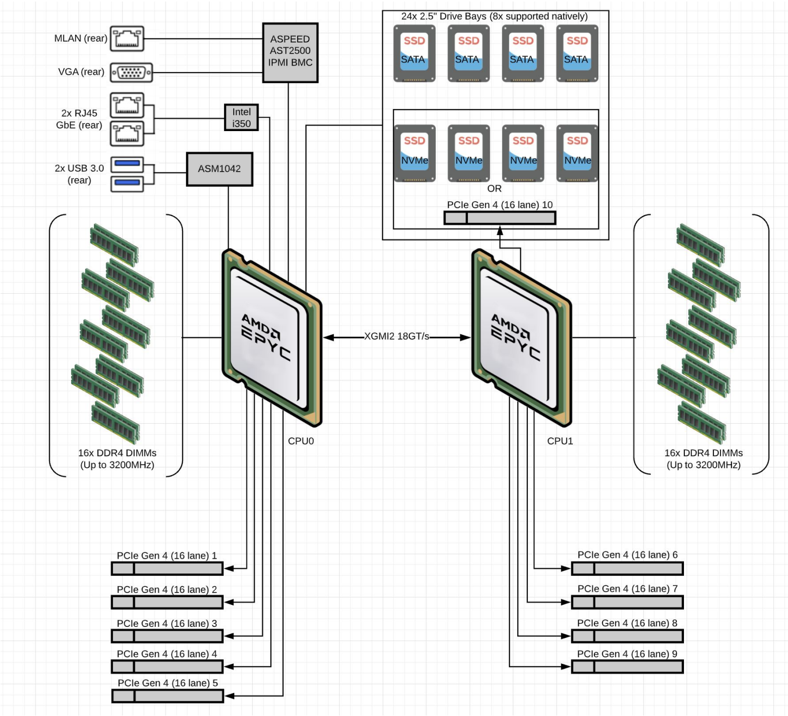
Approximately doubles the peak node-to-node bandwidth vs. double copy.



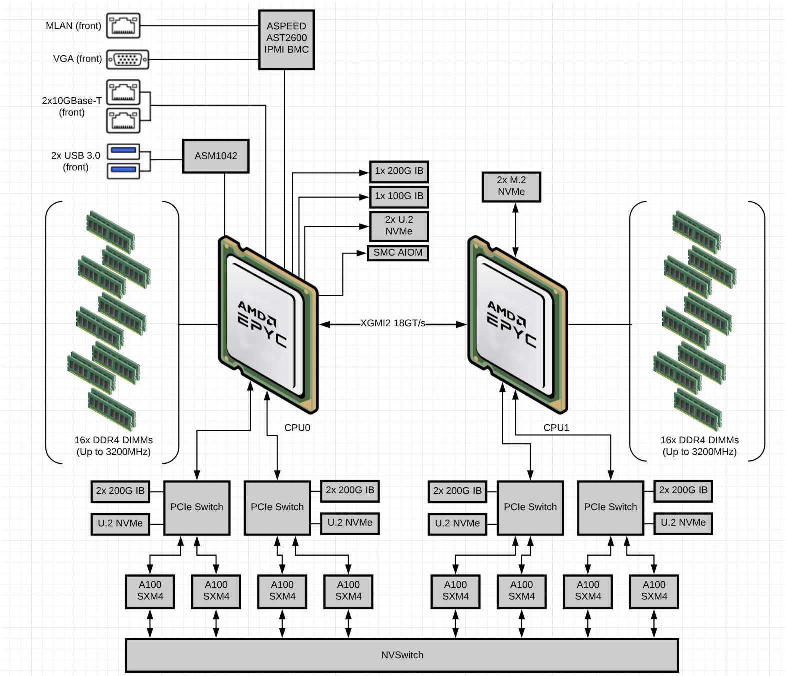
Data pathway without RDMA
(Additional copy to CPU memory)

Real life examples

Lambda Scalar GPU server PCIe topology



Lambda Hyperplane-8 A100 GPU server PCIe topology



Putting it all together

- ❏ Write your Cluster, Rack, and Node BOMs.
- ❏ Price your BOMs.
- ❏ Negotiate a colocation contract.
- ❏ Order all of the parts on your BOMs.
- ❏ Assemble the servers.
- ❏ Ship the racks / servers to the colo / DC.
- ❏ Rack, stack, label and cable the servers.
- ❏ Install your software.
- ❏ Turn it on.
- ❏ Maintain it.

Or, work with Lambda to build a customized Lambda Echelon GPU cluster for your team.

You’re done!

Citations

- Balaban, Sohmers, et al. **Lambda Echelon Deep Learning GPU Cluster - Reference Design Whitepaper.** https://lambdalabs.com
- Sergeev, Alexander and Mike Del Balso. **Horovod: fast and easy distributed deep learning in TensorFlow.** https://arxiv.org/pdf/1802.05799.pdf
- Pitch Patarasuk and Xin Yuan. **Bandwidth optimal all-reduce algorithms for clusters of workstations.** J. Parallel Distrib. Comput., 69:117–124, 2009. https://www.cs.fsu.edu/~xyuan/paper/09jpdc.pdf
- Jeaugey, Sylvain. **NCCL 2.0.** (2017). http://on-demand.gputechconf.com/gtc/2017/presentation/s7155-jeaugey-nccl.pdf
- WikiChip NVLink. https://fuse.wikichip.org/news/1224/a-look-at-nvidias-nvlink-interconnect-and-the-nvswitch/
- Mellanox Cable Management Guidelines. https://www.mellanox.com/sites/default/files/products/interconnect/pdf/Mellanox_Cable_Management_Guidelines_and_FAQs_Application_Note.pdf
- Berkeley Lab's report on ASHRAE Thermal Guidelines. https://datacenters.lbl.gov/sites/all/files/ASHRAE%20Thermal%20Guidelines_%20SVLG%202015.pdf
- Mellanox CS7520 Dismantling Guide: https://www.mellanox.com/related-docs/prod_ib_switch_systems/CS7520_Dismantling_Guide.pdf

Additional thanks to Chuan Li, Steve Clarkson, and Thomas Sohmers for assistance with this document.

About me

- CEO of Lambda.
- Started using CNNs for face recognition in 2012.
- First employee at Perceptio. We developed image recognition CNNs that ran locally on the iPhone. Acquired by Apple in 2015.
- Research published in SPIE and NeurIPS.

About **Lambda**

Lambda provides AI infrastructure to the Fortune 500, major research institutions, and the DOD.

Our products include the Lambda Quad workstation, Lambda Hyperplane server, **Echelon GPU cluster**, and the Lambda GPU cloud.

www.lambdalabs.com