WATGPU: A Hyperconverged GPU Research Platform

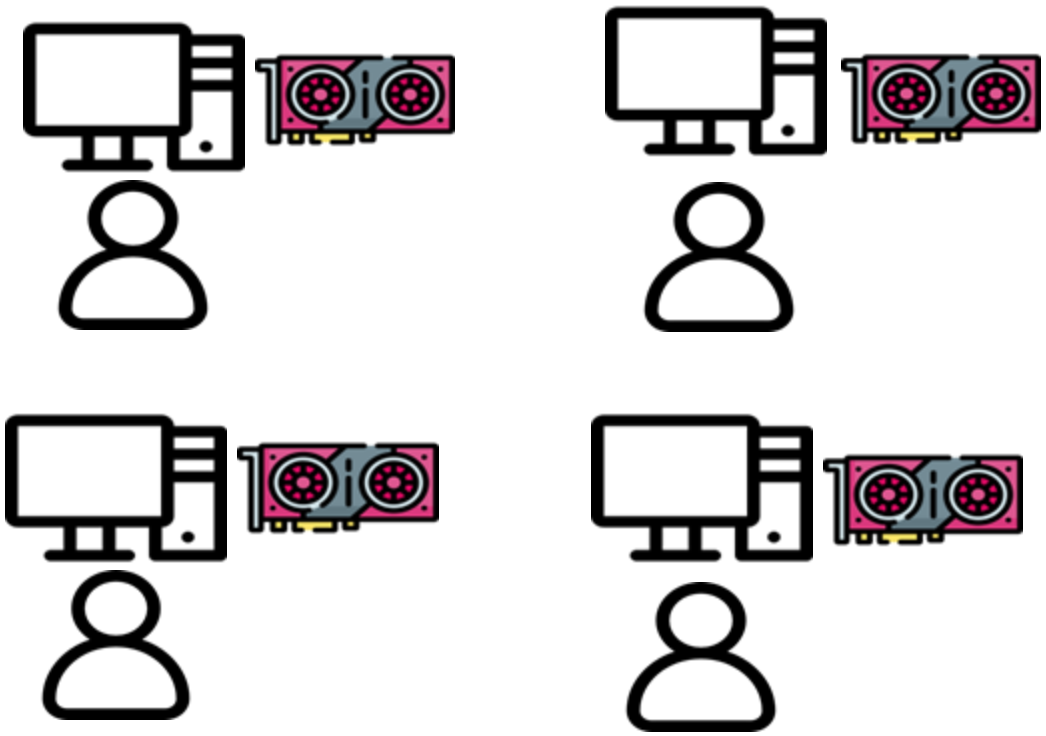
2024-12-02

Indy Ng (wk5ng@uwaterloo.ca)
Lucas Gamez (lmgamez@uwaterloo.ca)
Lori Paniak (ldpaniak@uwaterloo.ca)

Research and Special Projects (RSG)
Computer Science Computing Facility (CSCF)
Cheriton School of Computer Science

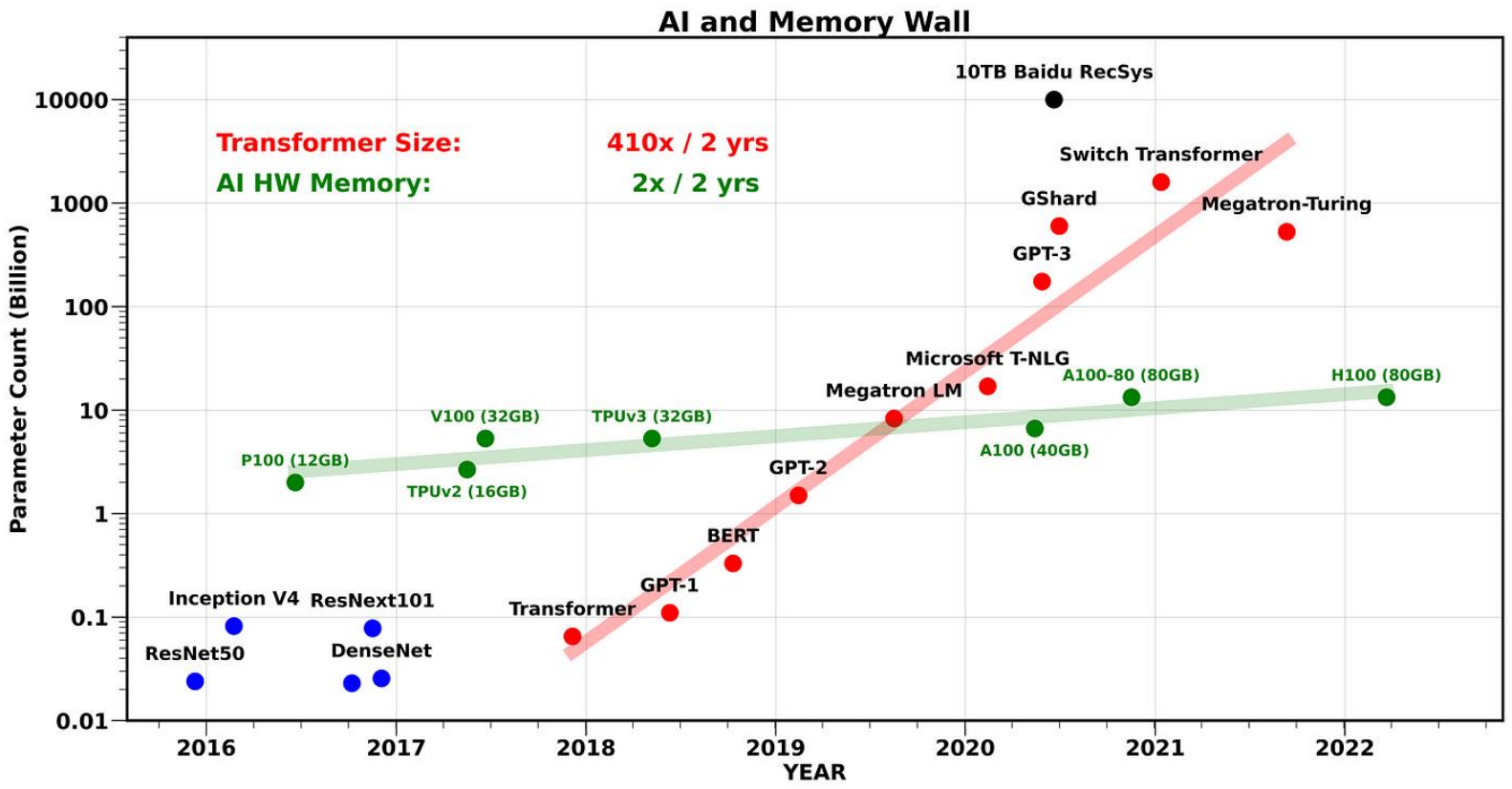
SOME BACKGROUND: What problem are we solving?

- New faculty get personal startup budget for research
- Want to turn that money into citations: buy relevant computer hardware



What are the problems with this situation?

- At best, 50/50 budget split on workstation and GPUs
- Workstations can be loud, hot, draw too much power
- GPUs only work for you, not active all the time
- Consumer GPUs vs datacenter GPUs



Note the logarithmic scale on the y-axis!

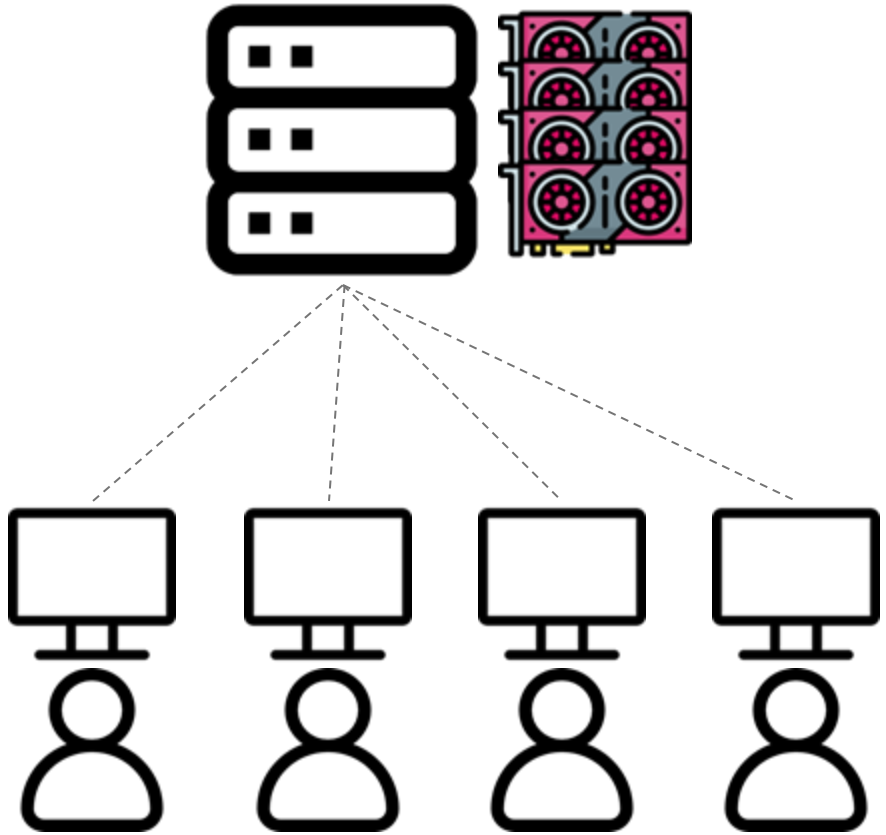
Consumer GPUs have limited VRAM (best is the RTX 4090 with 24GB)

Datacenter GPUs can do better, more VRAM and more TFLOPS!

Biggest GPUs do not have active cooling (fans)!
- depend on a server to push air

WHAT IS WATGPU?

- HPC servers that host School and faculty GPUs (GPU hotel)
- Hosts higher-end datacenter GPUs with more VRAM and TFLOPS (A6000, Ada 6000, L40S)
- Benefits of this setup:
  - Researchers provide the GPUs, school provides the servers
  - Central management, shared resources
  - Avoid idle GPUs and everyone buying expensive workstations/their own servers



Looks Cloudy: Why not Cloud?

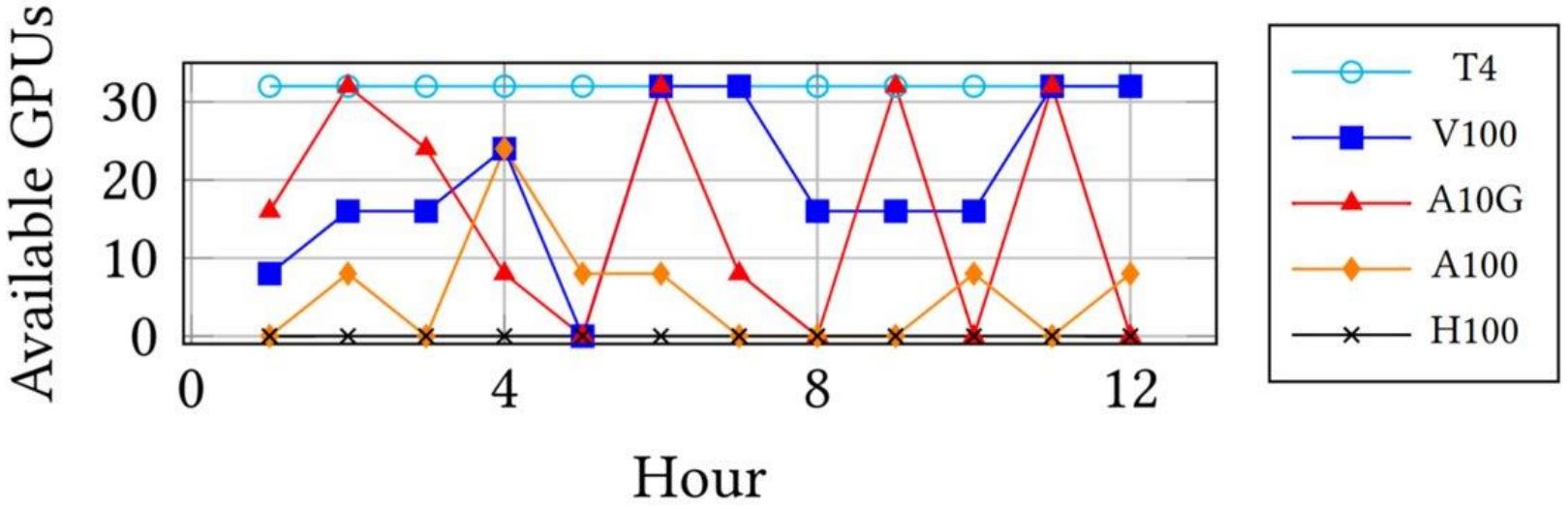
- nScale analysis: 8x A100 NVIDIA GPUs with substantial CPU/RAM https://vault.cs.uwaterloo.ca/s/cbdPbt76yQsMyib

  - Azure ND A100 v4 series Cloud long-term reserved access: **$478,000USD**
  - Buy the same system outright: **$182,000USD**

- Upgrade paths
- Can you even get high-end GPUs?

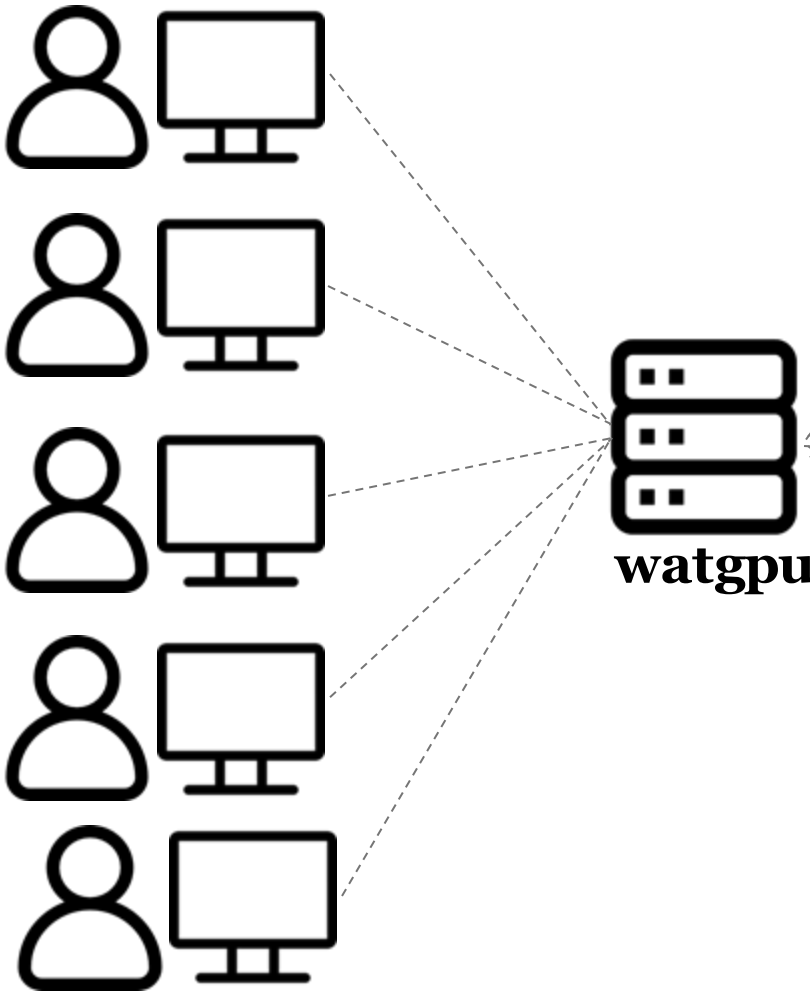
AWS GPU availability

- From 2024-11-30 seminar by Benson Guo: https://docs.google.com/presentation/d/1SJlw2gtZSKMtQ3sgCDJOOCm-LVc5rBiu8xXGNJjjUDw/edit?usp=sharing



UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS

WHAT DOES WATGPU LOOK LIKE?



watgpu

watgpu108
9 Ada 6000 GPUs
917 GB RAM
32 CPU cores

watgpu208
8 Ada 6000 GPUs
917 GB RAM
32 CPU cores

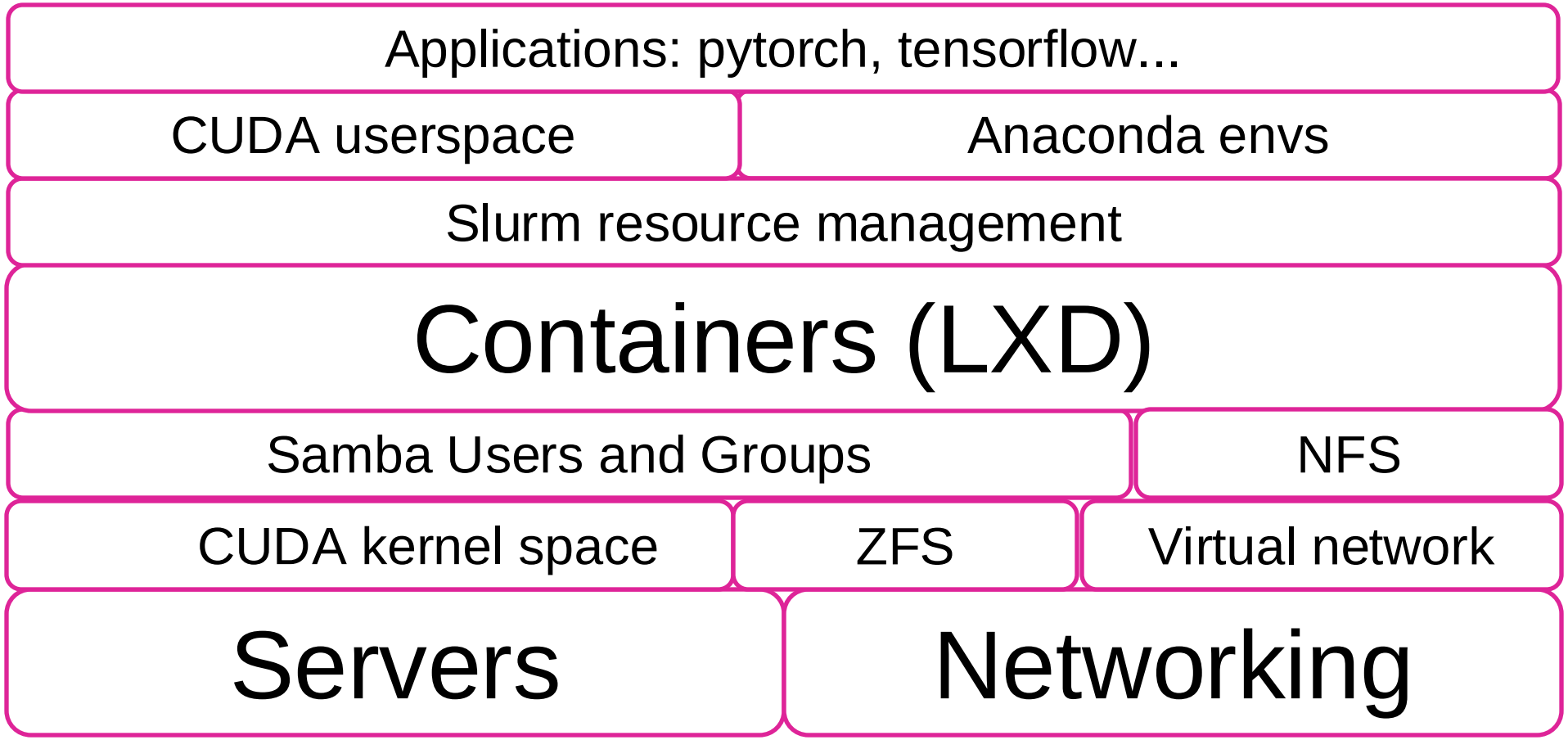
watgpu308
4 L40S, 2 Ada 6000, 2 A6000 GPUs
917 GB RAM
32 CPU cores

watgpu408
8 L40S GPUs
1160 GB RAM
128 CPU cores

Cerio chassis
ext-left: 4 Ada 6000 GPUs
ext-right: 2 Ada 6000, 2 A6000 GPUs

- 27 of the 41 GPUs are faculty-owned
- 107 unique users and 18,264 jobs since January 2024
- URAs, grad students, PhD students from at least 28 different faculty

The Layer Cake

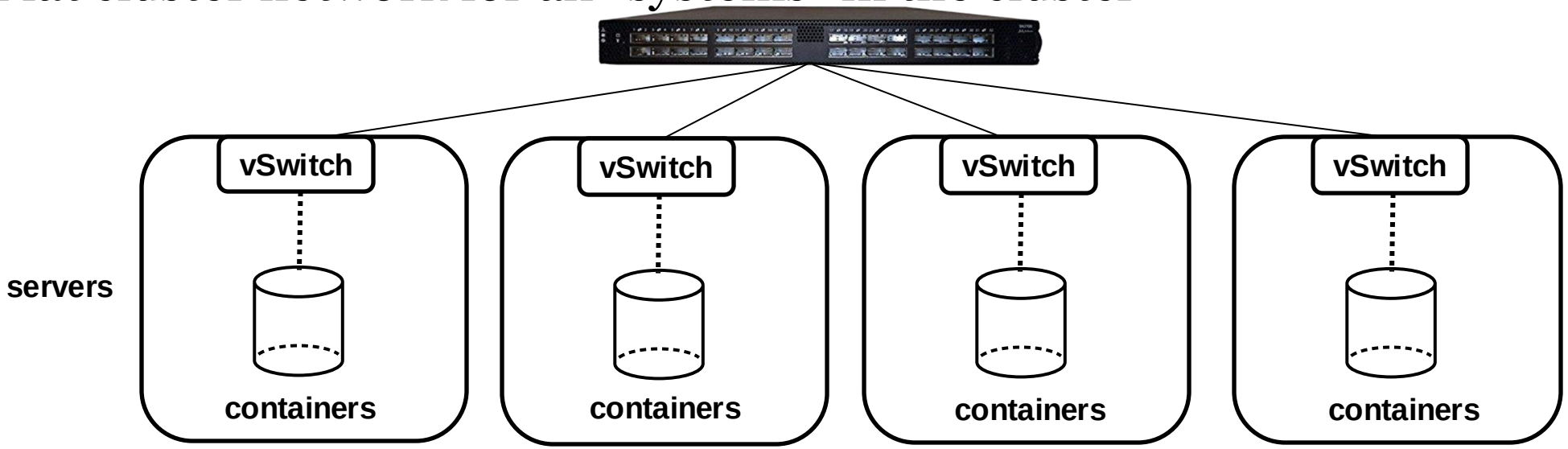


Servers

- Current server hardware consists of two types of high-capacity GPU servers:
  - Supermicro SYS-421GE-TNRT https://www.supermicro.com/en/products/system/gpu/4u/sys-421ge-tnrt
  - Supermicro AS -4125GS-TNRT1 https://www.supermicro.com/en/products/system/gpu/4u/as%20-4125gs-tnrt1
- Upcoming servers
  - Supermicro AS -5126GS-TNRT2 https://www.supermicro.com/en/products/system/gpu/5u/as%20-5126gs-tnrt2

Networking

- One 32-port 100Gbit switch
- Mellanox CX-6 Dx 100Gbit NICs on each machine (OCP cards, V5 PCIe x16)
- Virtual networks per server with Linux bridge connection to physical adapter
- Flat cluster network for all "systems" in the cluster



UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS

GPUs, GPUs, GPUs...

- Each server has 8-10 GPUs directly attached
  - NVIDIA RTX A6000: active cooling, 48GB VRAM, Ampere-class
    https://www.nvidia.com/en-us/design-visualization/rtx-a6000/
  - NVIDIA RTX 6000 Ada: active cooling, 48GB VRAM, Ada Lovelace-class
    https://www.nvidia.com/en-us/design-visualization/rtx-6000/
  - NVIDIA L40s: passive cooling, 48GB VRAM, Ada Lovelace-class
    https://www.nvidia.com/en-us/data-center/l40s/
- Drivers (only) for GPUs are installed on server OS: no userspace

CERIO EXPANSION CHASSIS

- Not all GPUs in the cluster are in servers!
- Disaggregated computing via Cerio chassis and PCIe bridges
- Hot-plug GPU capability
- https://www.cerio.io/cerio-platform/

```
|       +-03.4-[4b-50]----00.0-[4c-50]--+-00.0-[4d]----00.0  Cerio Placeholder Device
|       |                               +-01.0-[4e]----00.0  NVIDIA Corporation AD102GL [RTX 6000 Ada Generation]
|       |                               +-02.0-[4f]----00.0  Cerio Placeholder Device
|       |                               \-03.0-[50]----00.0  Cerio Placeholder Device
```

Storage: ZFS

- ZFS pool(s) on each physical machine
- Not redundant
- 7.68TB and 15.4TB NVMe devices backing each pool
- Each user gets their own filesystem
- Homes exported around the cluster via NFS
- Big data research == big data usage

```
root@watgpu-200:~# zfs list watgpu-200-u5-pool watgpu-200-pool
NAME                 USED  AVAIL     REFER  MOUNTPOINT
watgpu-200-pool     4.89T  1.77T       96K  none
watgpu-200-u5-pool  10.7T  3.11T       96K  none
```

USERS AND GROUPS

- Samba4 Active Directory for user account directory across the cluster
- Easy to set up and use
- Set up groups for various hardware contributor groups

CONTAINERS

- "Cluster" which is accessible to users are all LXD containers:
  - Login/gateway "machine"
  - Compute "machines"
- All connected via virtual networking within server and 100Gbit real networking between servers
- Applications can be modularly deployed in containers (XDMoD, JupyterHub):
  - Resource limits for each app
  - Independent app deployment
  - Choose what filesystems to mount, whether to domain join with Samba

SLURM

- Slurm was: Simple Linux Utility for Resource Management
- Resource manager for most TOP500 supercomputers
- A batch job submission system
- Various tools for tracking jobs:

```
root@watgpu:~# squeue |head
JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
28027       ALL test-cot    catai     PD               0:00        1               (Resources)
28136       ALL testspee   a9ding     PD               0:00      1               (Priority)
28086       ALL interact sszabado  R          2-17:10:56      1          watgpu208
```

SLURM

- sresources command to display resource availability

```
root@watgpu:~# sresources
NODE       STATE        GPUType                GPUTot     GPUAlloc   GPUFree        MEM        AllocMEM   FreeMEM        CPUTot     CPUAlloc   CPUFree
watgpu108  MIXED        ALL                    9          7          2              917G       893G       24G            128        50         78
                        | jimmygpu            |= 6       |= 5       |= 1
                        | schoolgpu            |= 1       |= 1       |= 0
                        | rcohengpu            |= 2       |= 1       |= 1

watgpu208  MIXED        ALL                    8          6          2              917G       885G       32G            32         12         20
                        | visiongpu            |= 8       |= 6       |= 2

watgpu308  IDLE+DRAIN   ALL                    8          0          8              917G       0G         917G           32         0          32
                        | yaolianggpu          |= 1       |= 0       |= 1
                        | schoolgpu            |= 7       |= 0       |= 7

watgpu408  MIXED        ALL                    8          7          1              926G       459G       467G           128        52         76
                        | jimmygpu            |= 4       |= 3       |= 1
                        | yaolianggpu          |= 3       |= 3       |= 0
                        | kfountouschoolgpu    |= 1       |= 1       |= 0
```

SLURM

- Create a (bash) script with details and send to Slurm:
  sbatch my-slurm-script.sh

```
#!/bin/bash
#SBATCH --time=00:05:00
#SBATCH --mem=4GB
#SBATCH --cpus-per-task=1
#SBATCH --gres=gpu:1
#SBATCH --partition=ALL

#SBATCH --job-name="Seminar example"
#SBATCH -o pytorch-test-run-%A.out
```

Lucas Gamez (lmgamez@uwaterloo.ca), DC 2607
Indy Ng (wk5ng@uwaterloo.ca), DC 2607
Lori Paniak (ldpaniak@uwaterloo.ca), DC 2625