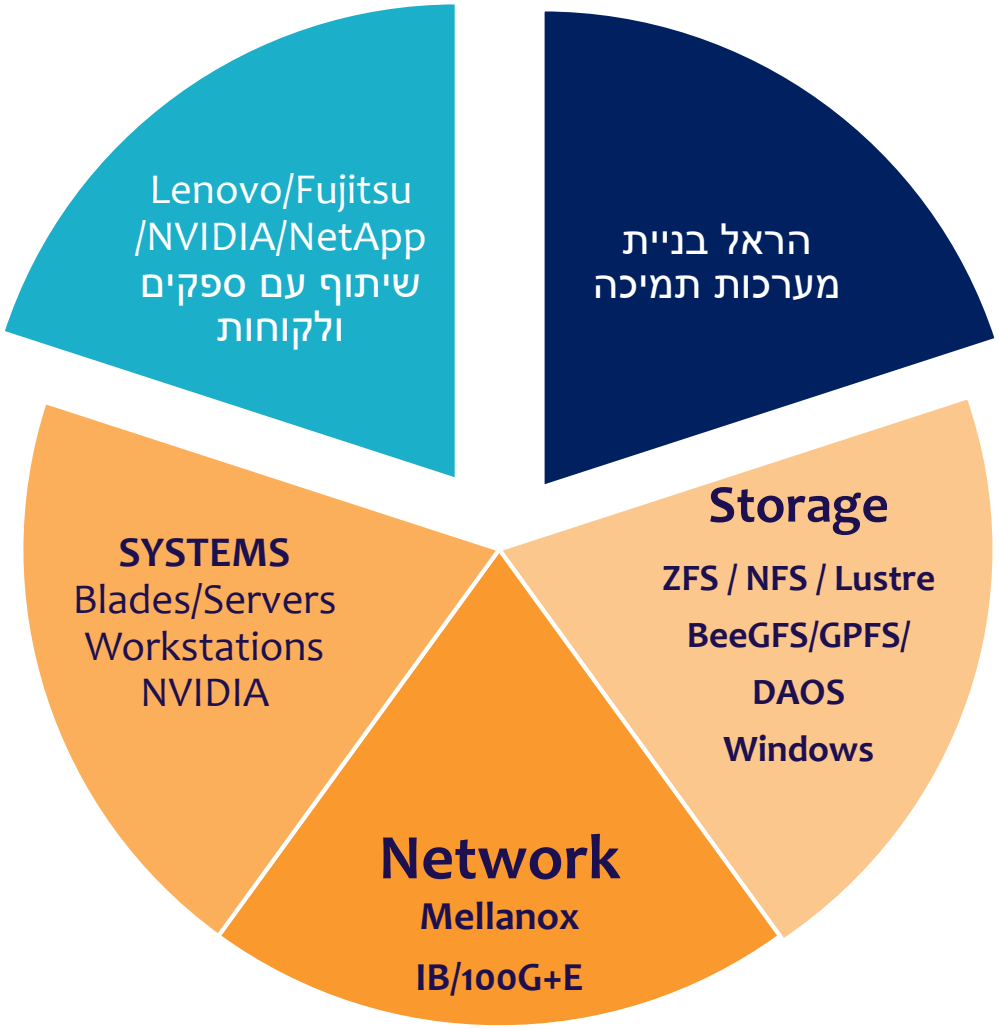
רקע למחשוב - על בהראל

בהנחיית:
מנהל תחום HPC הראל - אלכס לנדסברג

לינוקס, HPC, מערכות אחסון גדולות ומהירות, HPC משולב ענן, DEEP LEARNING, BIG DATÁ


משרד ראש הממשלה
ממשל זמין gov www.gov.il
משרד האוצר
משרד החינוך
משרד הבריאות
המרכז למיפוי ישראל SURVEY OF ISRAEL
משרד התמ"ת
רשות האוכלוסין וההגירה
משרד הבטחון
רשות ההגבלים העסקיים
משרד ההסברה והתפוצות
משרד התרבות והספורט
המשרד לשיוויון חברתי
המשרד לשירותי דת
משרד העלייה והקליטה
הלשכה המרכזית לסטטיסטיקה Central Bureau of Statistics دائرة الإحصاء المركزية
משרד המדע, הטכנולוגיה והחלל Ministry of Science, Technology & Space
רשות מקרקעי ישראל
המשרד לביטחון הפנים
משרד התחבורה והבטיחות בדרכים
משרד החקלאות ופיתוח הכפר
משרד התשתיות הלאומיות, האנרגיה והמים www.energy.gov.il
נציבות שירות המדינה
משרד המשפטים
משרד הפנים
המשרד לפיתוח הנגב והגלי מרחב האפשרויות של ישו
המשרד להגנת הסביבה
ישראל משרד התיירות
סלקום
IAI
אלביט מערכות מודיעין וסייבר
RAFAEL
אוניברסיטת חיפה UNIVERSITY OF HAIFA
האוניברסיטה העברית בירושלים
אוניברסיטת בן-גוריון בנגב
מכון ויצמן למדע
משרד התקשורת מדינת ישראל
משרד מבקר המדינה ונציבות תלונות הציבור
הפניקס
הבורסה לניירות ערך
הבינלאומי
ישראכרט
מזרחי טפחות
בנק הפועלים להיות ראשון זה מרחייב!
Cal עושים את זה קל
בנק ישראל BANK OF ISRAEL
leumi לאומי
דיסקונט
מחשב לכל ילד
חברת החשמל
www.jerusalem.muni.il שווה לחיות בירושלים
באר שבע בירת ההדמוניות של ישראל
אם נרצה הרצליה עיריית הרצליה
פתח תקוה עצב של עיר לב של מחשבה
TEL AVIV YAFO תל אביב יפו
כללית 100 שנה
מאוחדת אוהב אנשים
לאומית לנו זה אכפת

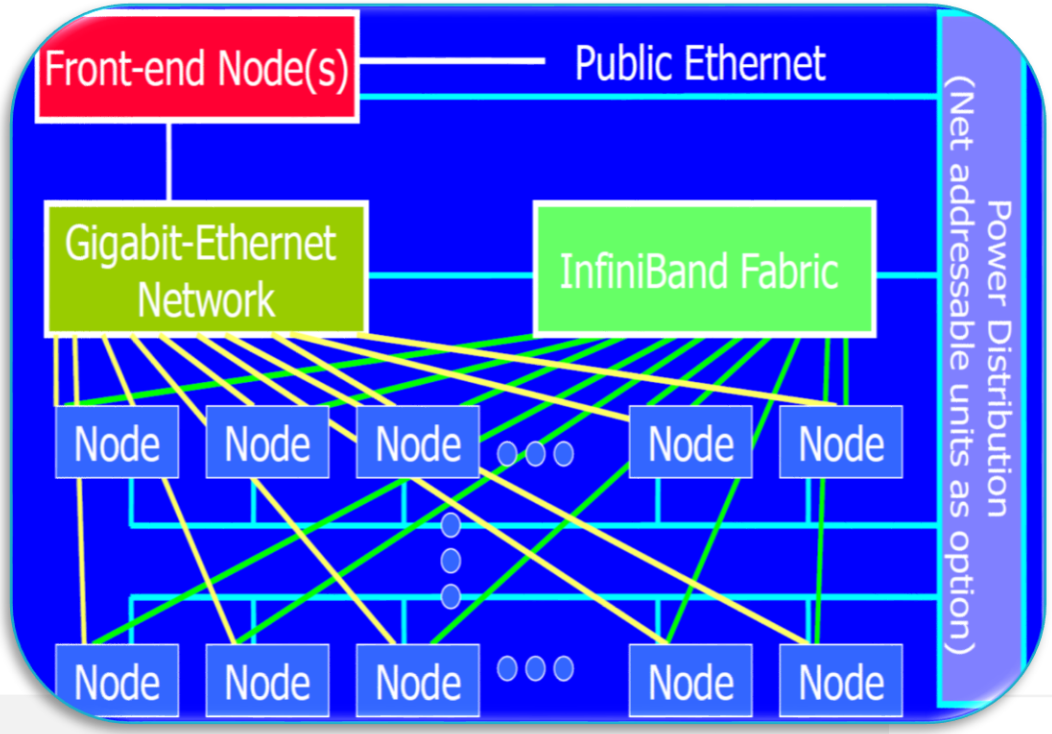
The Essence of Super Computing at Harel



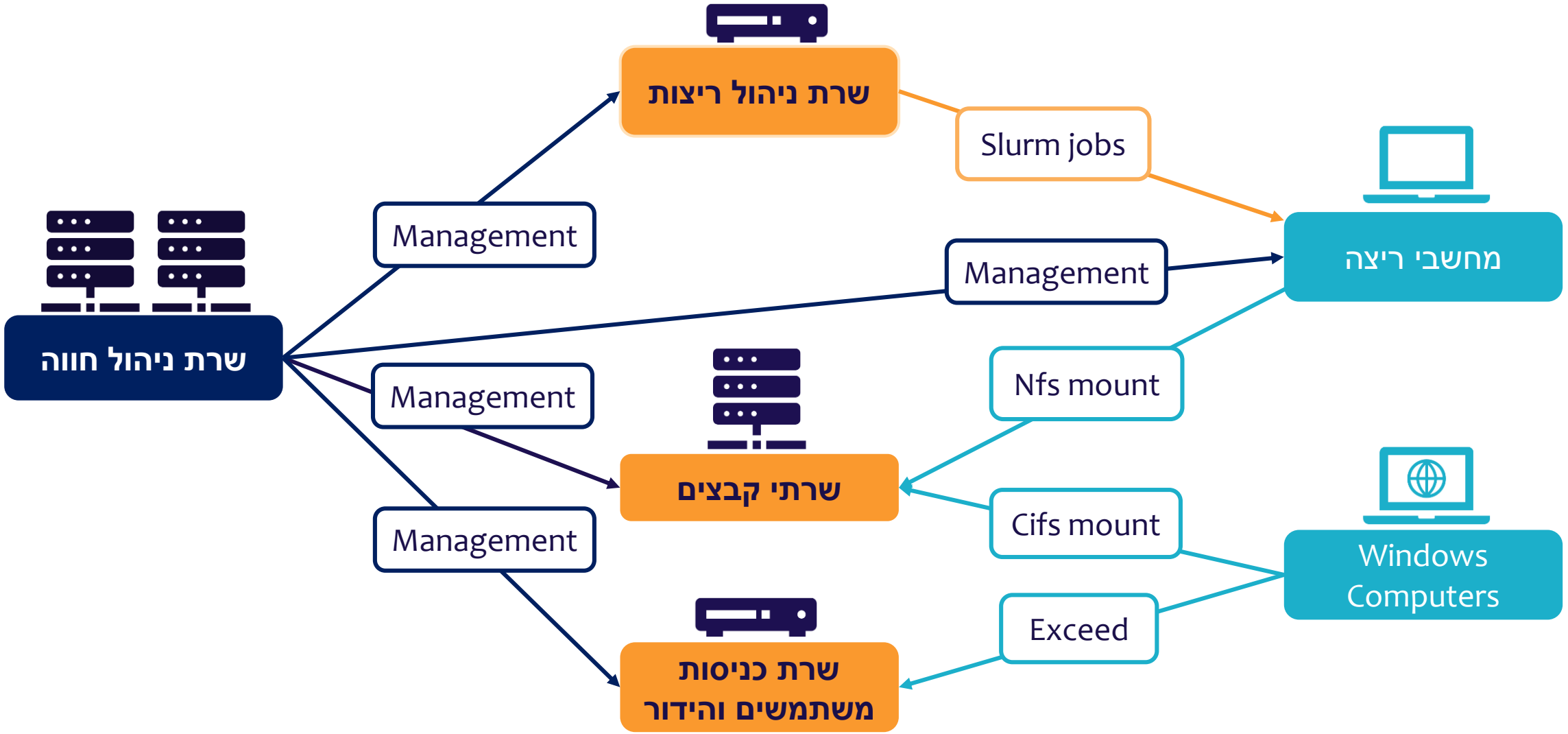
בחירת טכנולוגיות עבור מערכת HPC (High Performance Computing)

מבוסס על ניסיון של מערכות קיימות:

- LLNL, Argonne, Indiana ,CMU ,TACC ,Cambridge
- אונ' העברית, טכניון, בר-אילן, וכו'
- וכל מפעלי מערכת הביטחון ומכוני מחקר



סכמה לוגית של מערכת HPC-AI



מרכיבים של Opensource במפעלים ביטחוניים

Open Source Components

- Redhat	-> CentOS**, Roce Linux, Ubuntu**
- Bright/SGI Cluster Manager	-> OSCAR/xCAT/OpenHPC
- Batch Manager	-> PBS/**SLURM** - Condor
- **VMWARE**	-> **KVM**
- Local Graphics Accelerator	-> TightVNC with OpenGL support, x2go
- Monitoring software	-> Ganglia, Nagios, Cacti, Zabbix, **Prometheus, Grafana**
- Development	-> ddd, gdb, eclipse
- Editing	-> xemacs, pine, pico
- Matlab	-> Octave, scilab, scipy, matlabmpi
- Graphics Display	-> VisIt, Gnuplot
- Fluent/StarCD/Xflow	-> **OpenFOAM, SU2**
- Storage	-> **ZFS, Lustre, BeeGFS**, DAOS

Saves on the need for licenses

Software Stack

10/40/56/100/200/400 Gbit Ethernet/IB Networking

- Authorisation
- NFS (Client)
- NIS, NTP Server
- Prometheus
- Grafana
- xCAT
- Modules
- Compiler

Management

- SLURM/PBSpro
- MySQL
- Prometheus
- IB/Ethernet
- Linux

PBSpro/SLURM Server

- Backup Client
- ZFS
- Lustre/BeeGFS
- Prometheus
- IB/Ethernet
- Linux

Storage Servers

- Storage Client
- ZFS/Lustre/BGFS
- SLURM/PBS
- MPI
- Prometheus
- IB/Ethernet
- Linux

Compute Nodes

Bare Metal / Vmware / KVM / Containers

Software Stack

10/40/56/100/200/400 Gbit Ethernet/IB Networking

Authorisation
NFS (Client)
NIS, NTP Server
Prometheus
Grafana
xCAT
Modules
Compiler

Management

SLURM
MySQL
Prometheus
Ethernet
Linux

SLURM Server

Backup Client
NFS Server
Prometheus
Ethernet
Linux

Storage Servers

Applications
NFS
SLURM
MPI
Prometheus
IB/Ethernet
Linux

Compute Nodes

Bare Metal / Vmware / KVM / Containers

Products We Install

Applications

OpenFOAM,Xflow, Matlab,LS-DYNA, Numeca, Fluent, NES, MGAERO, SciLAB, ScaleMP
NAMD, Amber,GAMESS, CHARMM, Abaqus, ANSYS, CFD++, CFX, WRF,mitGCM, StarCCM+

Corporate Network

SLURM/
PBS

Login
Server

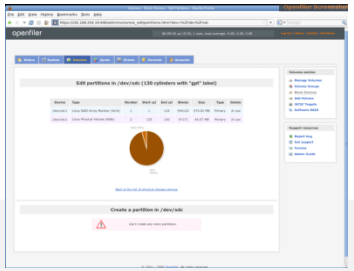
LUSTRE /NFS
/CIFS /RDMA

KVM/
VMware

GPU

Compute
Servers

ScaleMP



Mellanox
100/200Gb/sec

Cluster Monitoring

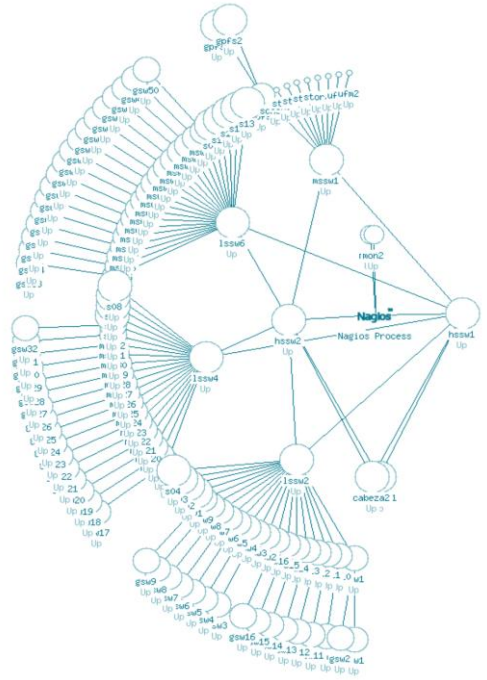
Cluster Monitoring

Nagios

Health
Monitoring

Nagios is an open source monitoring system which checks the availability of your HW and SW resources, notifies users of outages and generates performance data for reporting.

Scalable and extensible, Nagios can monitor large, complex environments across multiple locations.

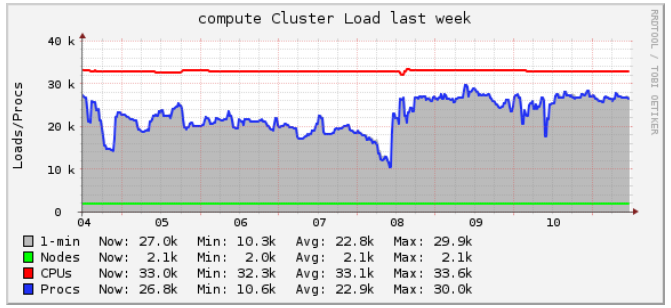


Ganglia

Performance
Monitoring

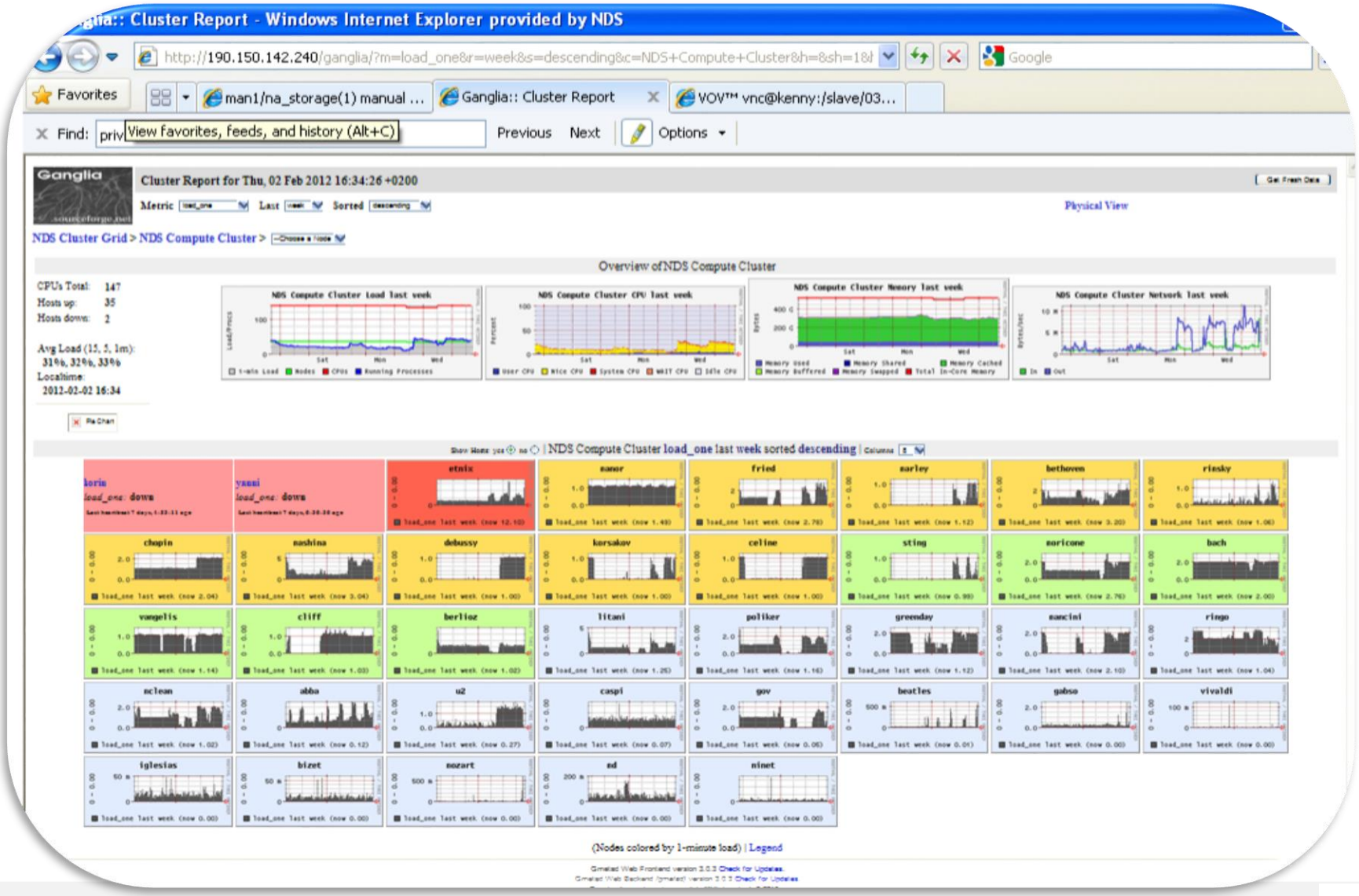
Ganglia is a scalable distributed monitoring system for high-performance computing systems such as clusters and Grids.

The implementation is **robust**, has very low per-node overheads and is currently **in use on thousands of clusters** around the world

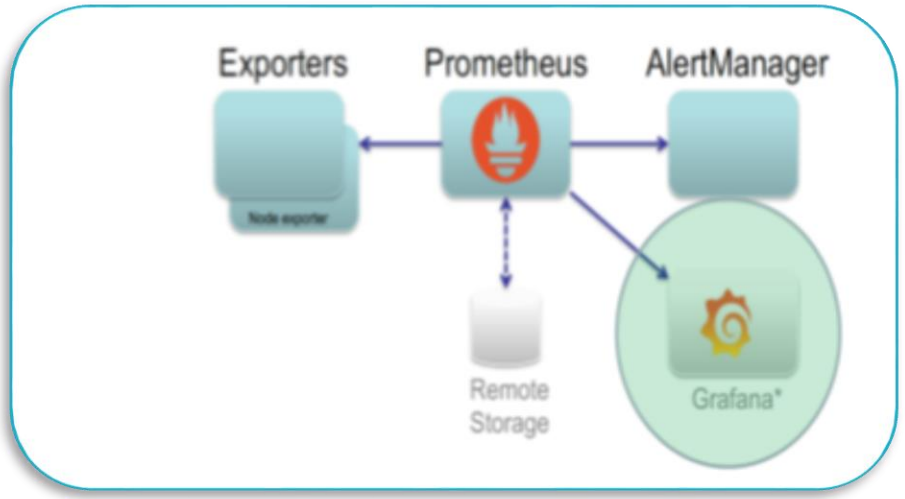


Monitoring

Ganglia Cluster Monitoring



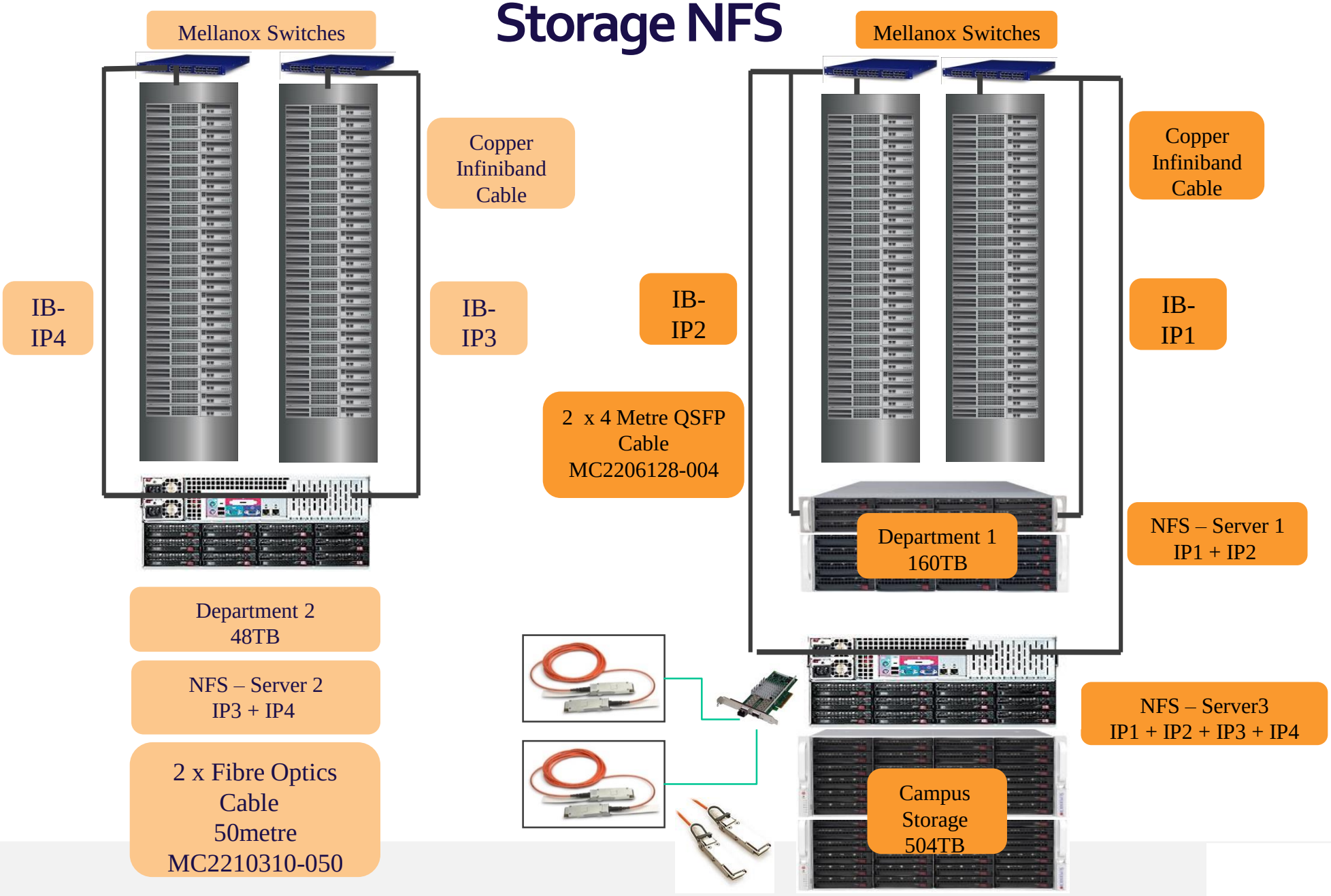
Prometheus Dashboards



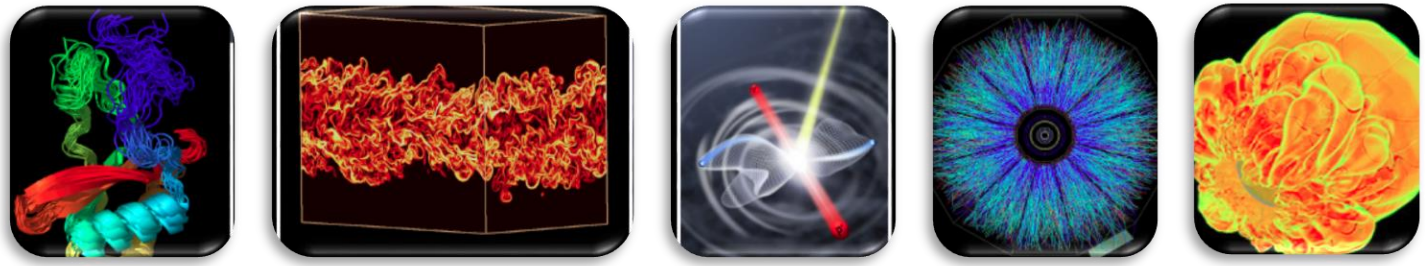
Grafana

Custom Dashboards

Storage NFS


Mellanox Switches
Copper Infiniband Cable
IB-IP4
IB-IP3
Department 2 48TB
NFS – Server 2 IP3 + IP4
2 x Fibre Optics Cable 50metre MC2210310-050
Mellanox Switches
Copper Infiniband Cable
IB-IP2
IB-IP1
2 x 4 Metre QSFP Cable MC2206128-004
Department 1 160TB
NFS – Server 1 IP1 + IP2
NFS – Server3 IP1 + IP2 + IP3 + IP4
Campus Storage 504TB

ZFS on
Standard Hardware

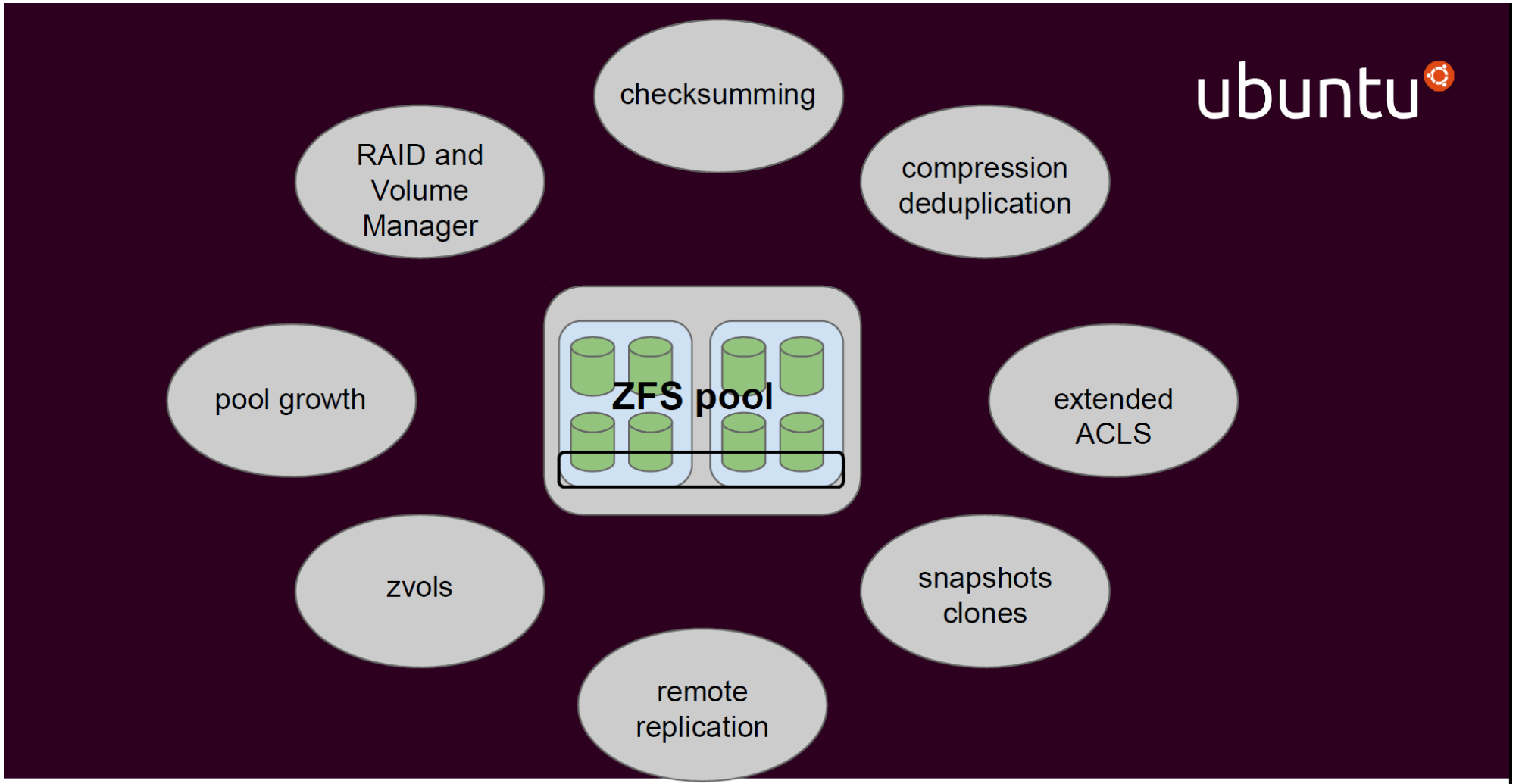


Trever Nightingale
Senior Systems Analyst
NERSC Server Team

ZFS State of the art

- 2^{48} – Number of entries in any in individual directory
- 16 EB – Maximum size of a single file
- 16 EB – Maximum size of any attribute
- 256 ZB (2^{78} bytes) - Maximum size of any zpool
- 2^{56} - Number of attribute of a file (constrained to 2^{48} for the number of files in a ZFS file system)
- 2^{64} - number of devices in any zpool
- 2^{64} - number of zpool in a system
- 2^{64} - Number of file systems in a zpool

ZFS – File System – Now Ubuntu Standard



Tutorial: How to install, tune and Monitor a ZFS based Lustre file system

2nd annual Lustre Ecosystem Workshop

Marc Stearman
Lustre Operations Lead

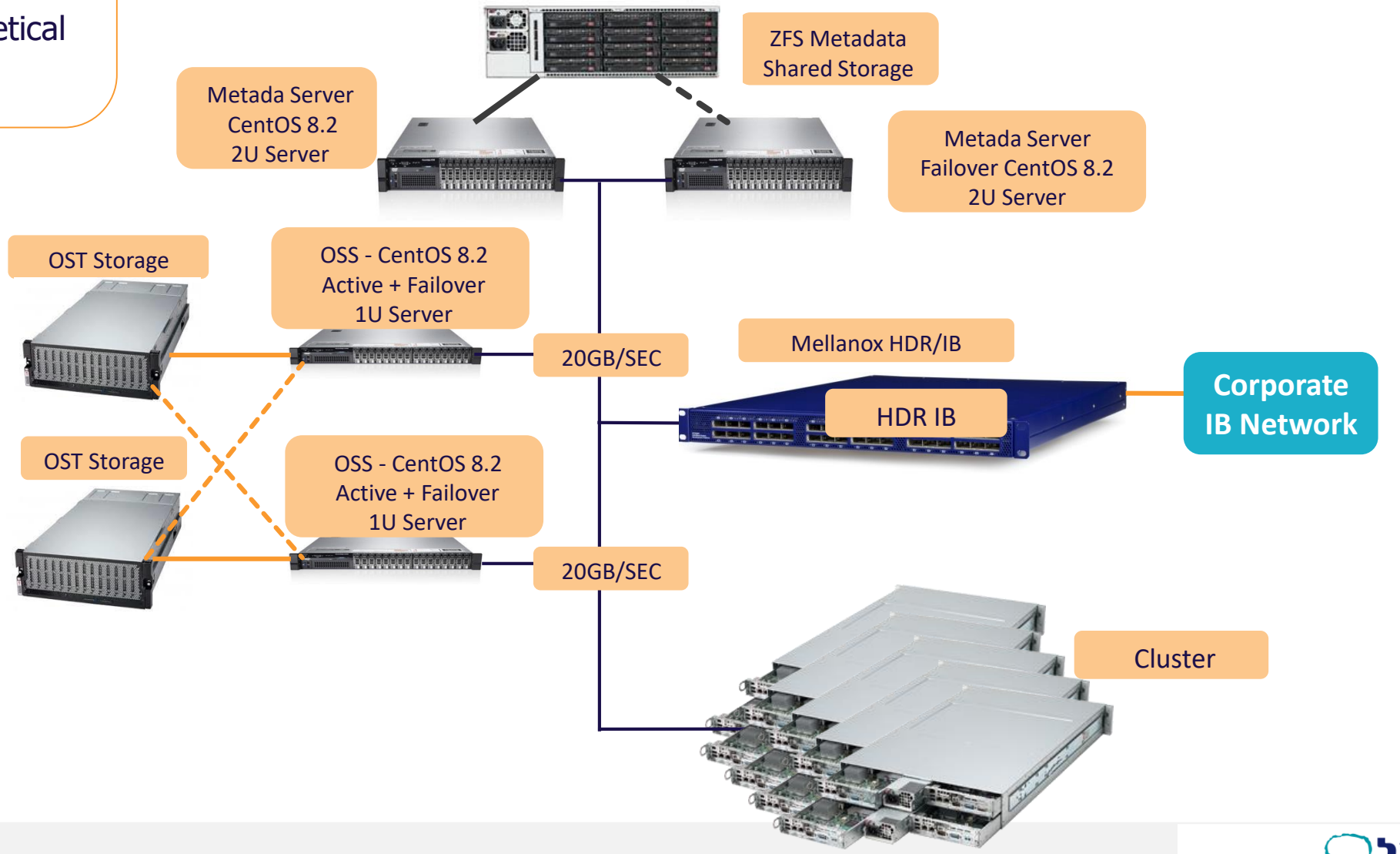
March 9-10, 2016

-PRES-683717
was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore
eratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

Lawrence Livermore National Laboratory

2x OSS – 2 MDTs/ZFS
Total 180 Disks * 6TB
>1PB Storage
HDR IB
20GB/Sec/OSS
40GB/Sec Total Theoretical Throughput
=~30GB/SEC

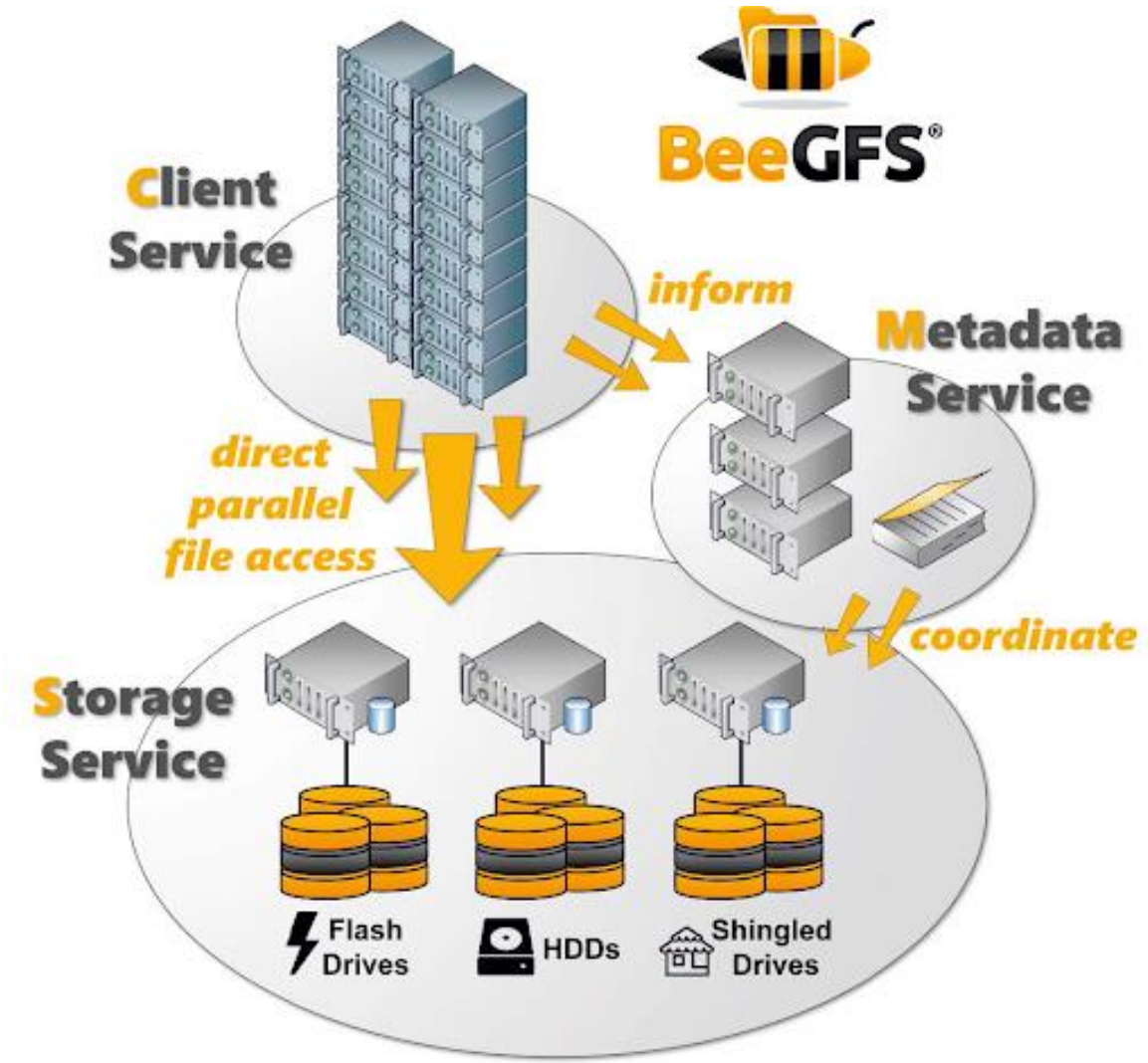
Parallel Storage - Lustre



ORCAM

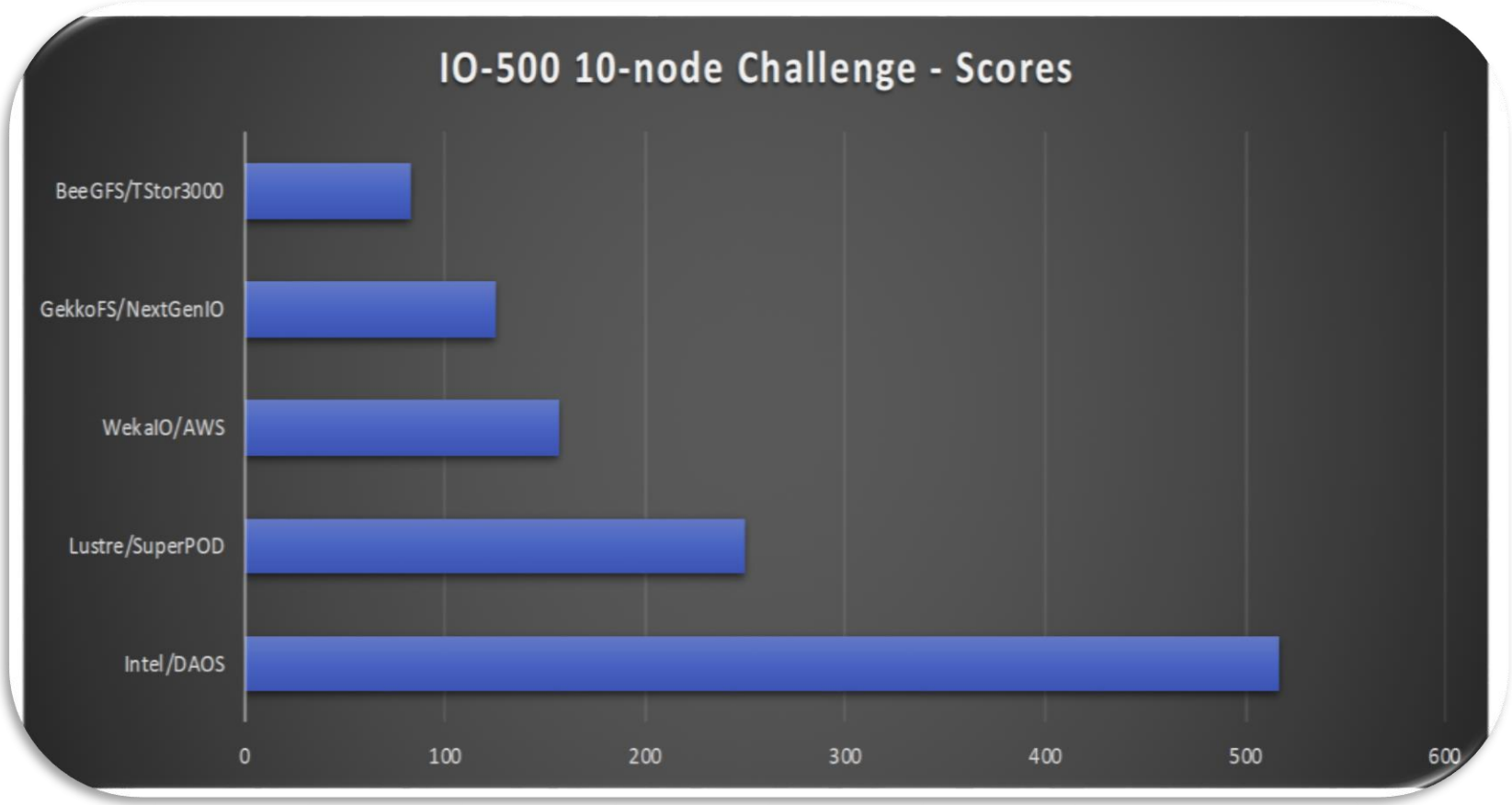
10 Servers
With SSD Disks
ZFS Fully Parallel FS
Dual 10GE

Parallel Storage - BeeGFS



Future Directions With Storage

DAOS & IO - 500 – 10 – node Challenge



Distributed Asynchronous Object Storage

DAOS: Primary Storage on Aurora

Aurora DAOS configuration

- Capacity: 230PB
- Bandwidth: >25TB/s

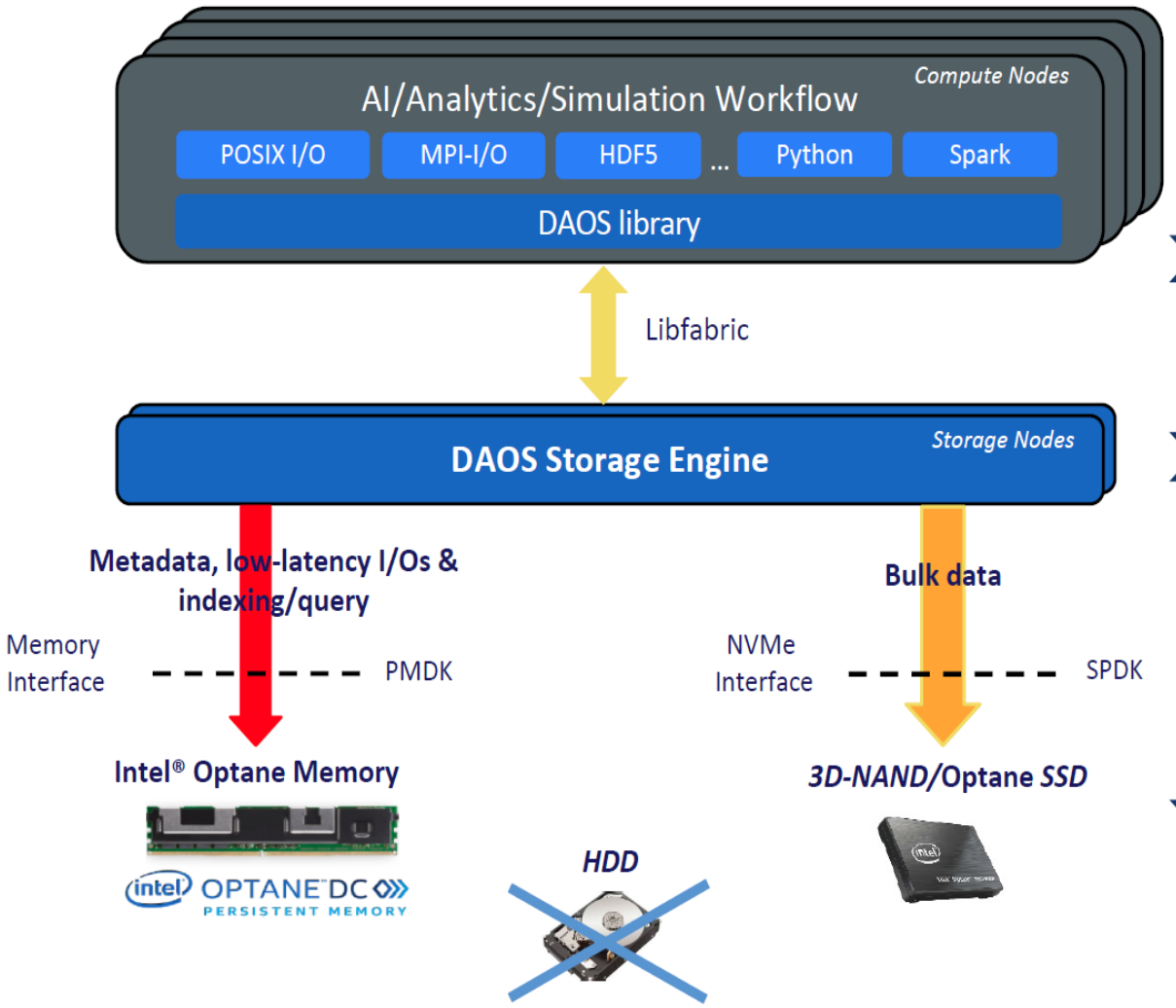
"Combined in Aurora, the Intel compute system, Cray Slingshot network, and the Intel DAOS storage open new possibilities for accelerating the scientific research needed to solve critical human challenges such as cancer and disease. DAOS enables the creation of new storage data models tailored specifically to applications like the Cancer Distributed Learning Environment (CANDLE) which provide a powerful platform to advance a wide array of scientific challenges using deep learning."

– Rick Stevens, Associate Laboratory Director for Computing, Environment and Life Sciences

"The Argonne Leadership Computing Facility is excited to be the first major production deployment of the DAOS storage system as part of Aurora, an US exascale system coming in 2021. As designed, it will provide us unprecedented levels of metadata operation rates and extremely high bandwidth for I/O intensive workloads."

– Susan Coghlan, ALCF-X Project Director/Exascale Computing Systems Deputy Director

DAOS Storage Architecture



- DAOS library directly **linked** with the applications
- No need for **dedicated** cores
- **Low** memory/CPU **footprint**
- End-to-end OS **bypass**
- Non-blocking, lockless, snapshot support, …

- **Low-latency** & **high-message-rate** communications
- Native support for **RDMA** & **scalable** collective operations
- Support for iWarp, RoCE, Infiniband, OPA, Slingshot, …

- **Fine-grained I/O** with media selection strategy
- Only application data on SSD to **maximize throughput**
- Small I/Os **aggregated** in pmem & migrated to SSD in large chunks
- Full userspace model with **no system calls** on I/O path
- **Built-in** storage management infrastructure (control plane)
- NFSv4-like **ACL**

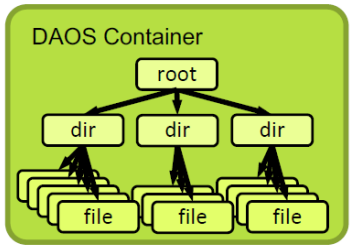
Delivers high-IOPS, high-bandwidth and low-latency storage with advanced features in a single tier

DAOS Storage Containers

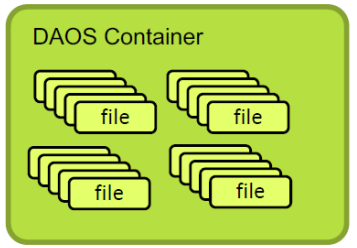
Storage Containers

Aggregate related datasets into manageable and coherent entities

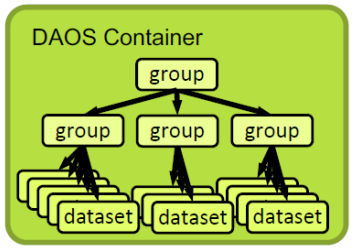
- Distributed consistency & automated recovery
- Full Versioning
- Simplified data management
  - Snapshot
  - Cross-tier Migration
  - Indexing



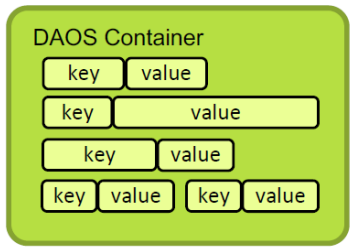
Encapsulated POSIX Namespace



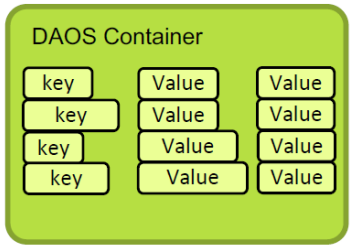
File-per-process



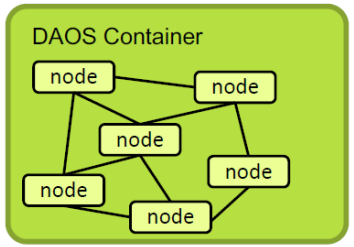
HDF5 « File »



Key-value store

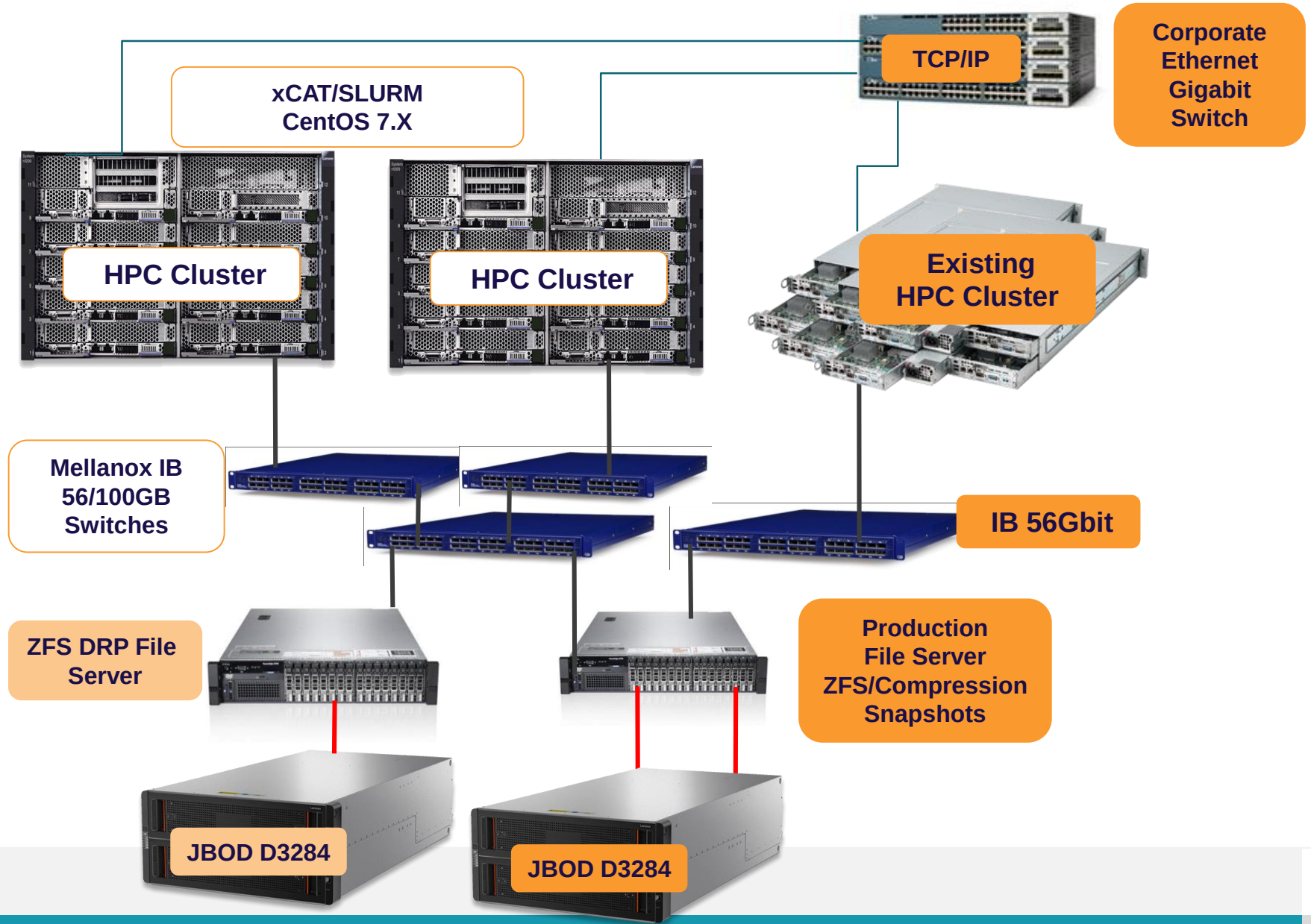


Columnar Database

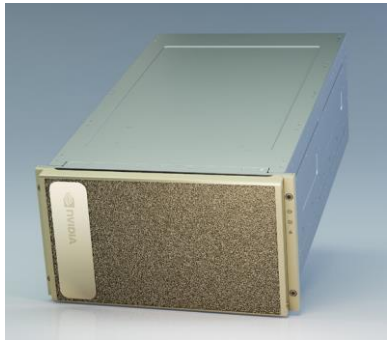


Graph

Example Cluster - OpenU



AI & NVIDIA SYSTEMS



NVIDIA
DGX 8*A100

Fujitsu GX2570 M6
8*A100

Lenovo SR670 V2
4XA100

NVIDIA ELITE PARTNER

FUJITSU SELECT Expert

Data Center Partner Platinum Lenovo

Harel-PS

HPC

- xCAT
- Infiniband Libraries
- Intel and GCC Compilers
- OpenMPI
- MPI
- SLURM/PBS

- MATLAB
- R
- OpenFOAM
- STARCCM+
- LSDYNA
- ANSYS
- Gaussian
- AMBER
- NAMD
- Abaqus
- WRF
- mitGCM

Deep Stuff

- cuDDN
- CUDA
- cuBLAS
- NCCL
- NVLink
- Hadoop
- Spark

- Python
- NVCaffe
- Caffe2
- Microsoft Cognitive Toolkit
- MXNet
- Tensorflow
- Theano
- PyTorch
- Torch
- DIGITS
- TensorRT

How do we Bridge Both 

Harel-PS

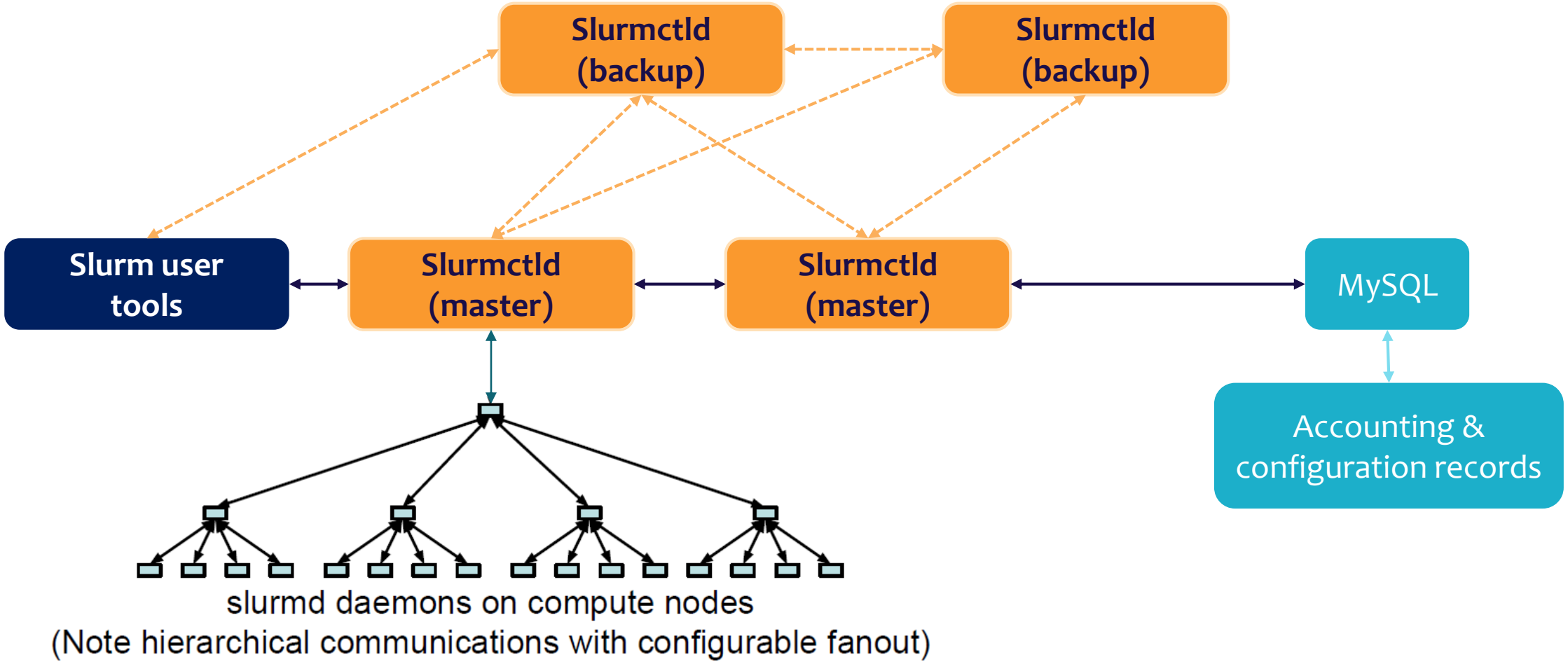
What's the Secret?
Containers

DOCKER,
THE MOST POPULAR CONTAINER


Sylabs.io
Home
Singularity
Docs
Resources
About Us
Cloud Login
SINGULARITY
SIMPLE, FAST, SECURE
LEARN MORE
THE CONTAINER PLATFORM FOR
PERFORMANCE SENSITIVE WORKLOADS
Enabling and securing your performance critical applications
from the core, through the cloud, and out to the edge.

SLURM Cluster Layout
Accounting and Billing

Israel Government
Defense Industries
Technion,HUJI,Haifa,BGU,OpenU, TAU
NVIDIA, LLNL, Argonne



Example of running a SLURM script program utilizing DGX systems

> srun -G 2 –pty dgx tensorflow
Use 2 GPUs, run it from an NVIDIA container with the latest version of tensorflow

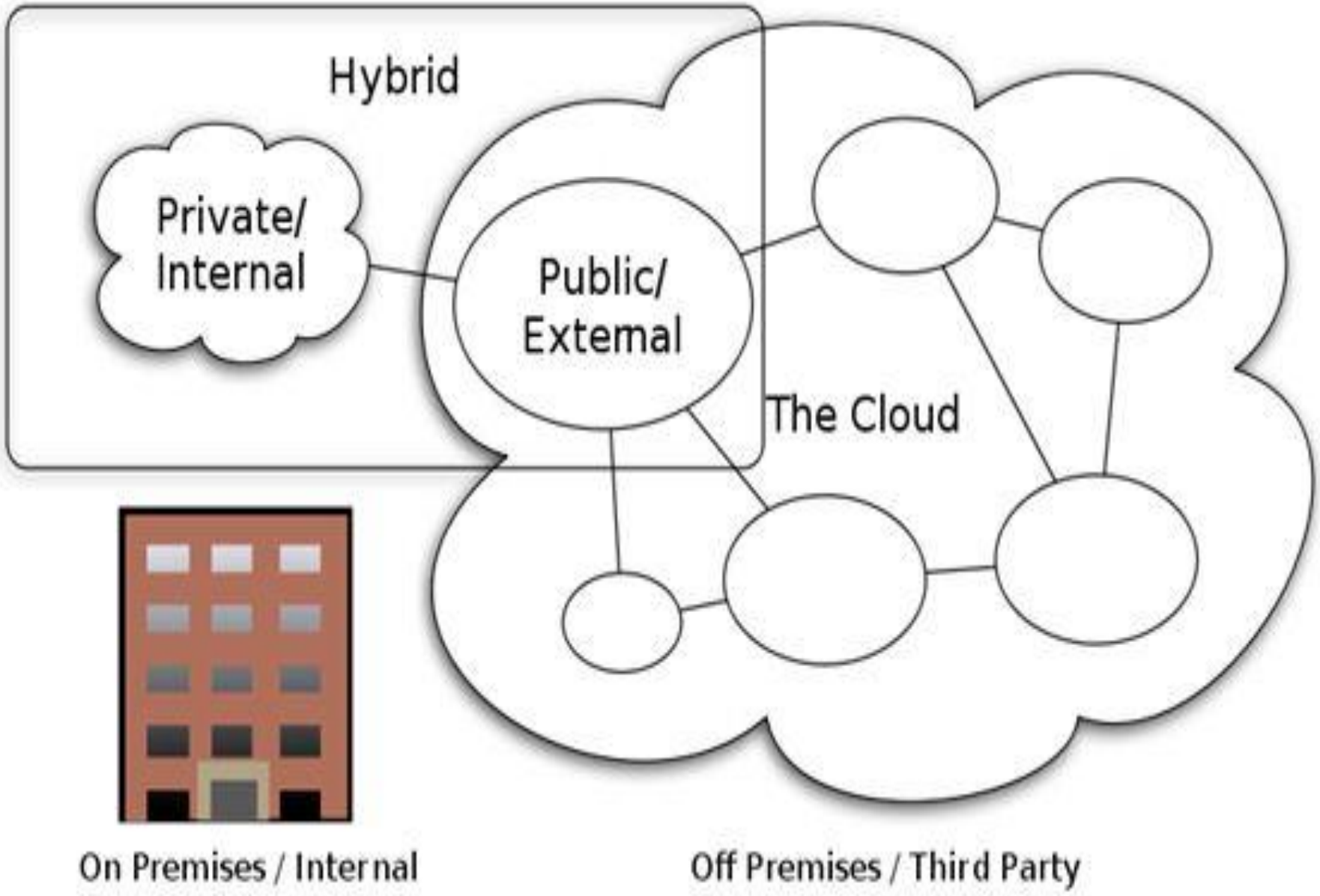
Example of SLURM batch script running on a DGX system

> cat example.script

```
#!/bin/bash -l
#SBATCH -p dgx
#SBATCH --gres=a100:1
#SBATCH --export=HOME,USER,TERM,WRKDIR
module load  nvidia-tensorflow
nvidia-docker run --rm -ti tensorflow/tensorflow \
```

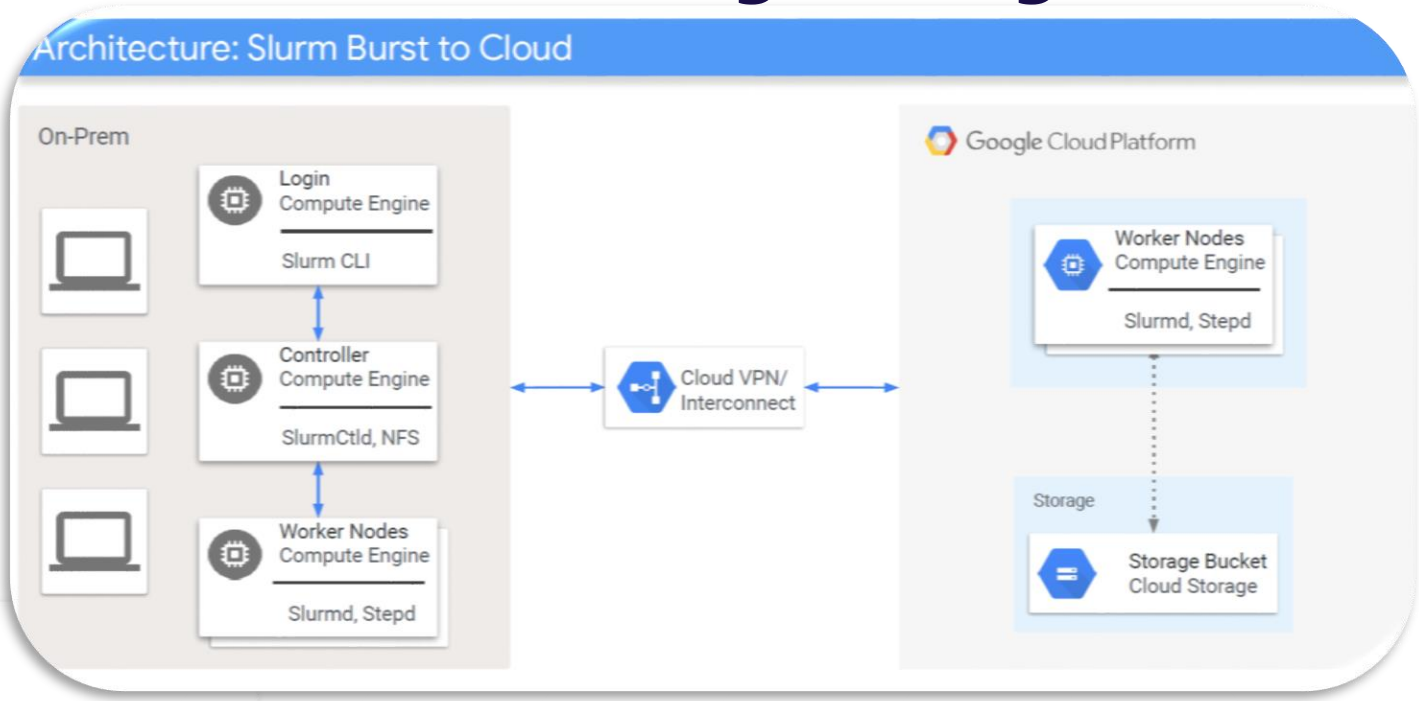
> sbatch example.script

Use partition partition dgx, and run on one A100.
Load the module nvidia-tensorflow and execute python from within a docker container.

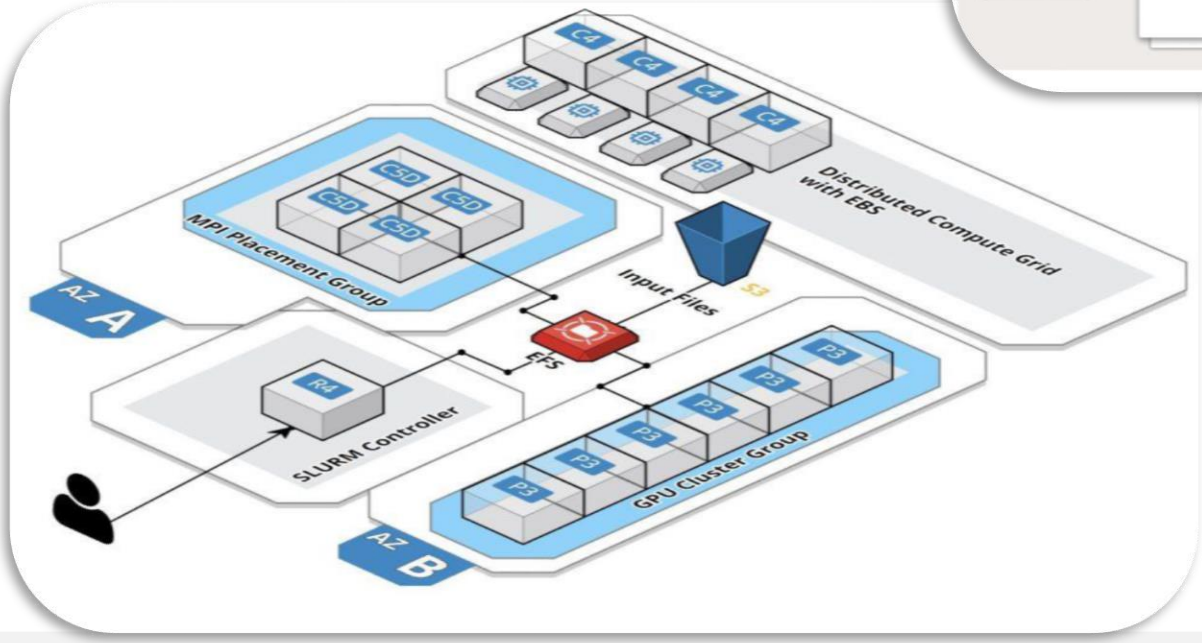


Cloud Computing Types

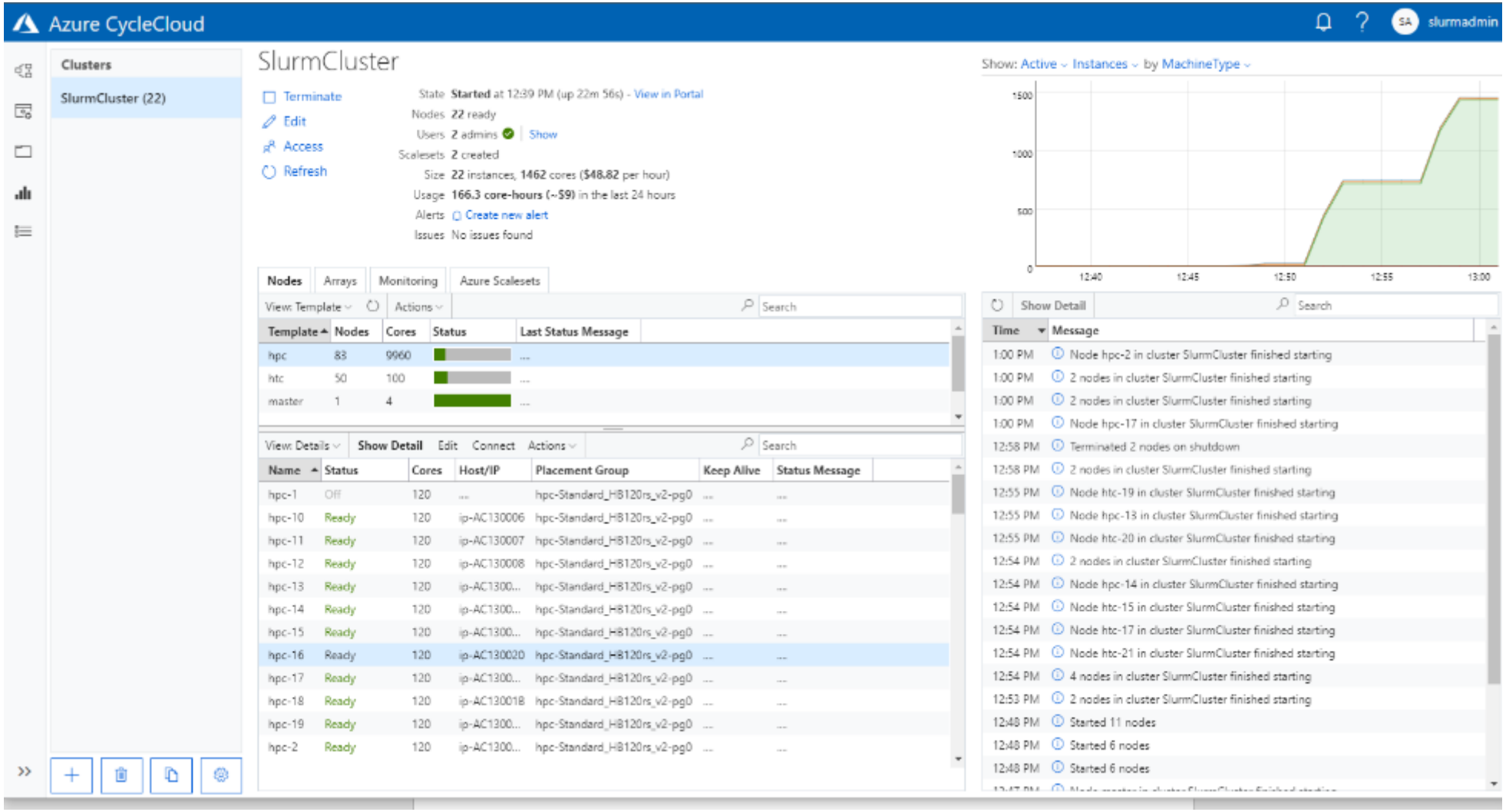
SLURM Bursting to Google Cloud



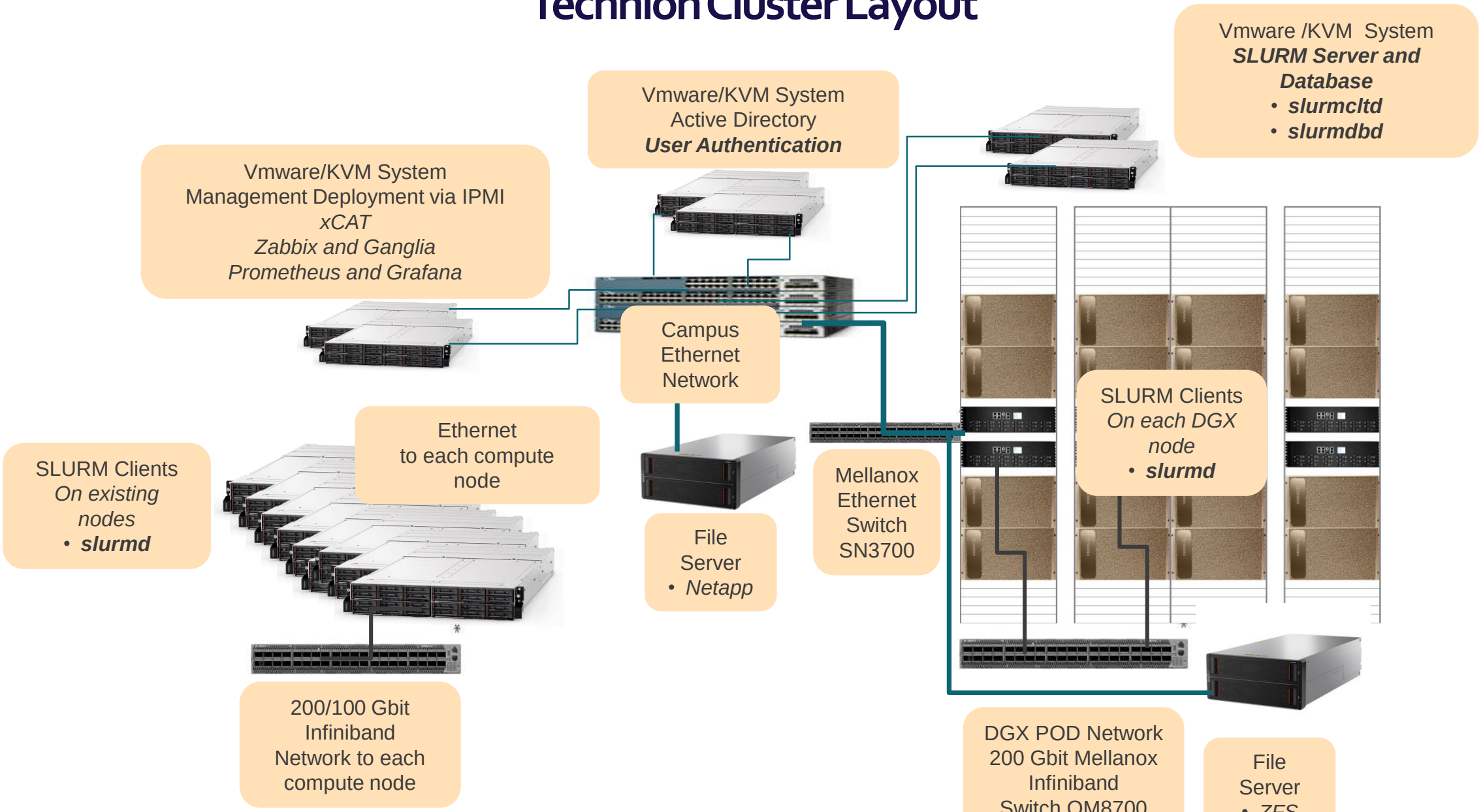
הרחבה ל- AWS מבוסס SLURM



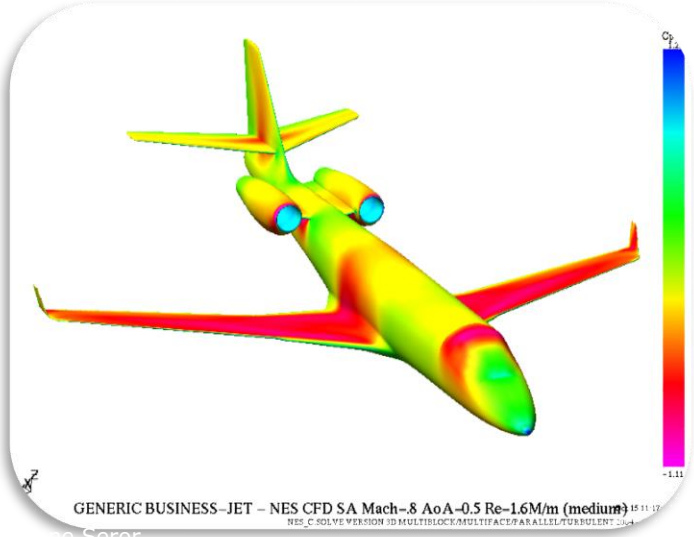
SLURM in Azure Cloud

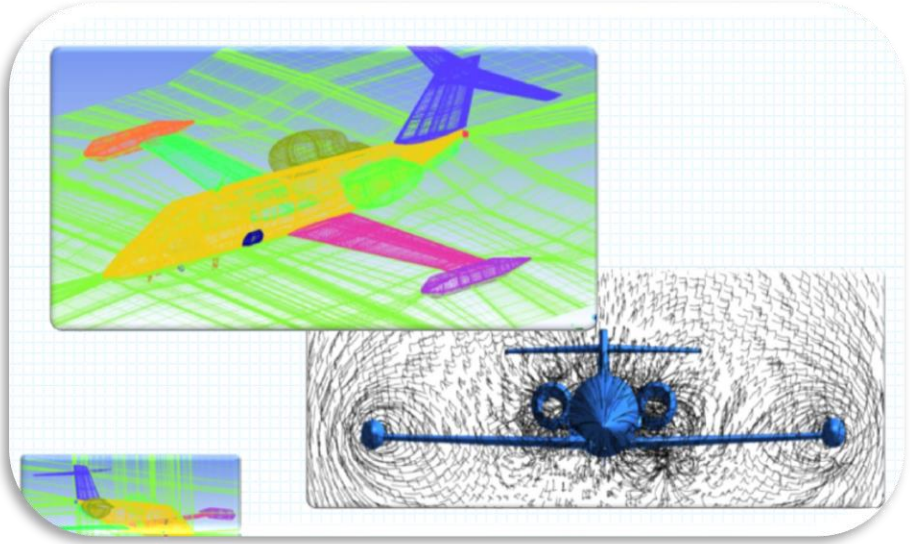


Technion Cluster Layout



Design using Open Source Cluster Software





(With courtesy of Stephane Seror
IAI Head Hypersonic R&D CFD branch)

Harel Supercomputing by Harel-PS
The Essence of Elegant Computing

PRESALES - אבינועם
avinoamz@harel.co.il

אביב - שירותים מקצועיים
Aviv@Harel.co.il

HPC & AI - אלכס
Alexl@Harel.co.il


הראל
טכנולוגיות מידע בע"מ
by one