Microsoft

# Performance Optimization of the CMAQ model on Microsoft Azure

Steve Roach

Azure / HPC Technical Specialist
sroach@microsoft.com

# Agenda

1. Describe Azure Platform
2. Discuss Factors which impact application performance
    a) Process Pinning
    b) Storage

# Azure: cloud built for HPC & AI

Genuine HPC & AI approach: platforms, benchmarks, people

Purpose-built hardware for the best performance, optimized price-performance and differentiated solutions

TTM availability of leading hardware innovations to accelerate "time to solutions" for customer workloads

Strategy to leverage leading internal production workloads using the same systems for mission critical offerings on Azure.

| Azure | Other clouds |
|---|---|
| Supercomputing scale for the most demanding applications | Azure is the only public cloud provider offering the full range of HPC and AI capabilities |
| InfiniBand HPC & AI clusters for best performance on real workloads | |
| Compute optimized VMs with low latency ethernet | Compute optimized VMs with low latency ethernet |

# Azure looks a lot like a HPC Datacenter

**A/B**

## A/B series VMs

Burstable virtual machines (VMs)

**D/E/F**

## General purpose VMs

D: Standard workloads

E: High memory

F: Compute bound

**L/M**

## Optimized GP VMs

L: High SSD & IOPS

M: Extreme memory

**H/N**

## Specialized VMs

HB/HC: CPU Supercomputing

ND: GPU Supercomputing

NC: GP-GPU compute

NV: Graphics / visualization

NP: Programmable FPGA

## HPE Cray in Azure

Shasta EX Supercomputers

Managed custom bare-metal server

Large to extreme-scale HPC infrastructure

Azure network integration

**Small scale MPI**
(Handful of cores)

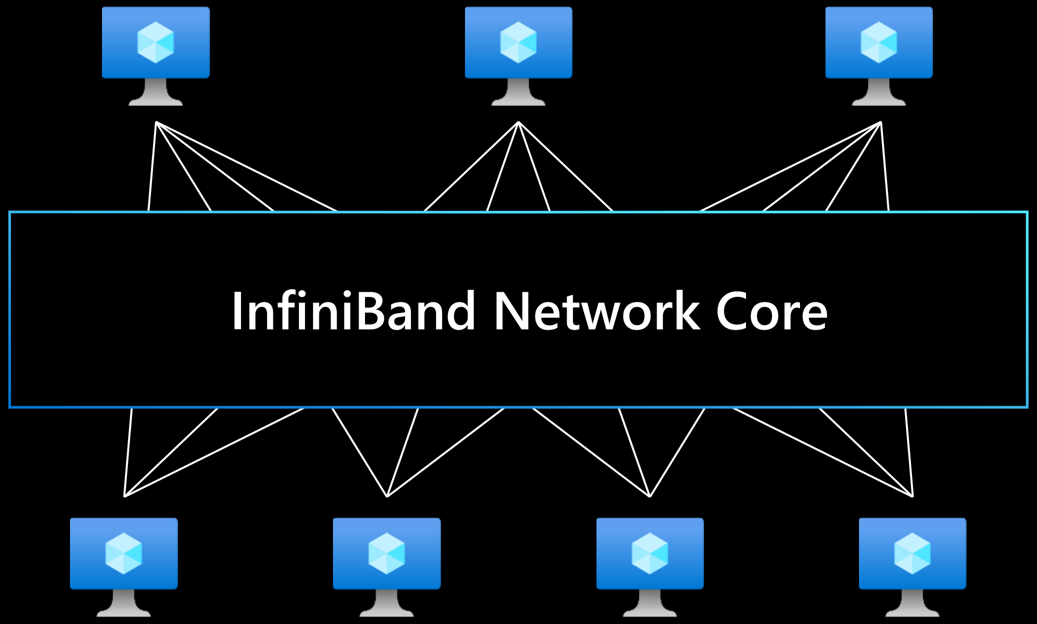
**Large to extreme scale MPI**
(100k+ cores)

Non-blocking Fat Tree topology

Hardware offload of MPI collectives

Full MPI & NCCL Integration

< 1.5 microsecond latencies




Up to 1.6 Tb/s per VM

Bare-metal passthrough

Intelligent Adaptive Routing

Dynamic Connected Transport

# SR-IOV Goodness on Azure

**MPICH Derivatives:**

- ✓ MPICH
- ✓ IntelMPI
- ✓ MVAPICH2
- ✓ Microsoft MPI
- …

**OpenMPI Derivatives:**

- ✓ OpenMPI
- ✓ HPC-X
- ✓ Platform MPI
- …

- Feature parity with bare metal
- Prior to SR-IOV enablement Supported only IntelMPI (dapl) and MSMPI

# HBv3 Upgrade

## Milan-X comes to Azure HPC

HBv3 Virtual Machines enhanced with AMD 3rd Gen EPYC with 3D v-cache

No customer/partner changes required, same HBv3 VM sizes

3x L3 cache increase per core, chiplet, socket, and VM (1.5 GB)

Accelerates HPC applications bound by memory performance

Increases *effective* memory bandwidth up to ~630 GB/s

Decreases *effective* memory latency by as much as 51%

Performance & Scalability of HBv3 VMs with Milan-X CPUs

# What is Milan-X and how does it affect performance?

Architecturally, Milan-X differs from Milan only by virtue of having 3x as much L3 cache memory per core, CCD, socket, and server.

| CPU | Xeon 2690 v4 "Broadwell" | Xeon Gold 6148 "Skylake" | Xeon 8280 "Cascade Lake" | EPYC 7742 "Rome" | EPYC 7V73X "Milan-X" |
|---|---|---|---|---|---|
| Cores/2S server | 28 | 40 | 56 | 128 | 128 |
| L3 cache/2S server | 70 MB | 55 MB | 77 MB | 512 MB | 1,536 |
| Relative size | 1x | 0.8x | 1.1x | 7.3x | 22x |

Examples of workloads that can benefit from larger L3 cache are:

- Computational fluid dynamics (CFD) – memory bandwidth
- Weather simulation – memory bandwidth
- Explicit finite element analysis (FEA) – memory bandwidth
- EDA RTL simulation – memory latency

# Advantages of Fewer Cores per Node

| | 120 Cores | 96 Cores |
|---|---|---|
| RAM per Core | 3.75 GB | 4.67 GB |
| Memory B/W per Core | 2.91 GB/s | 3.65 GB/s |
| L3 Cache per Core | 12.8 MB | 16 MB |

# Importance of Process Pinning

For MPI applications, optimal pinning of processes can lead to significant application performance improvements for under subscribed systems.

In the chiplet design, AMD has essentially integrated a bunch of smaller CPUs together to provide a socket with 64 cores (8 - 16 smaller CPUs with 4-8 cores each).

**To maximize the performance from each core it is important to balance the amount of L3 cache and memory bandwidth per core.**

Optimal MPI Process Placement for Azure HB Series VMs - Microsoft Community Hub

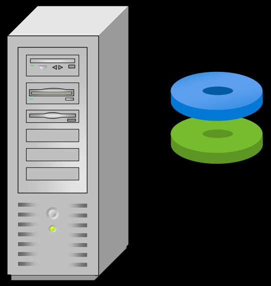
# Pinning Example

setenv PIN_PROCESSOR_LIST "--bind-to cpulist:ordered --cpu-set 0,1,2,3,4,5,8,9,10,11,12,13,16,17,18,19,20,21,24,25,26,27,28,29,30,31,32,33,34,35,38,39,40,41,42,43,46,47,48,49,50,51,54,55,56,57,58,59,60,61,62,63,64,65,68,69,70,71,72,75,76,77,78,79,80,81,84,85,86,87,88,89,90,91,92,93,94,95,98,99,100,101,102,103,106,107,108,109,110,111,114,115,116,117,118,119 --report-bindings "

( /usr/bin/time -p mpirun -np $NPROCS $PIN_PROCESSOR_LIST --rank-by slot -mca coll ^hcoll -x LD_LIBRARY_PATH -x PATH -x PWD $BLD/$EXEC ) |& tee buff_${EXECUTION_ID}.txt

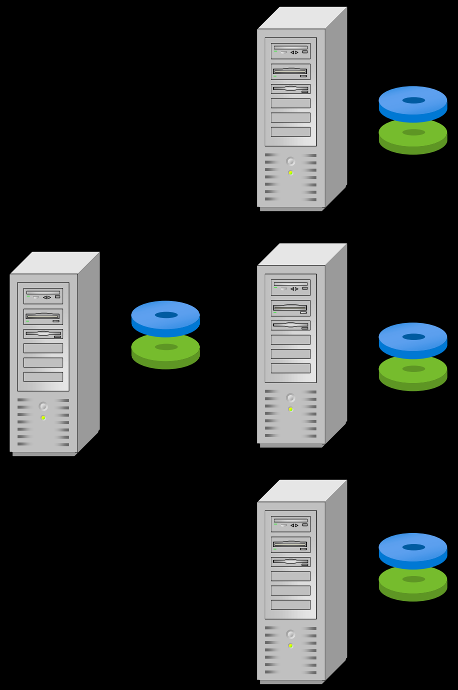
# Storage Options

Azure Files
with NFS



Stand Alone
NFS Server

Azure NetApp
Files



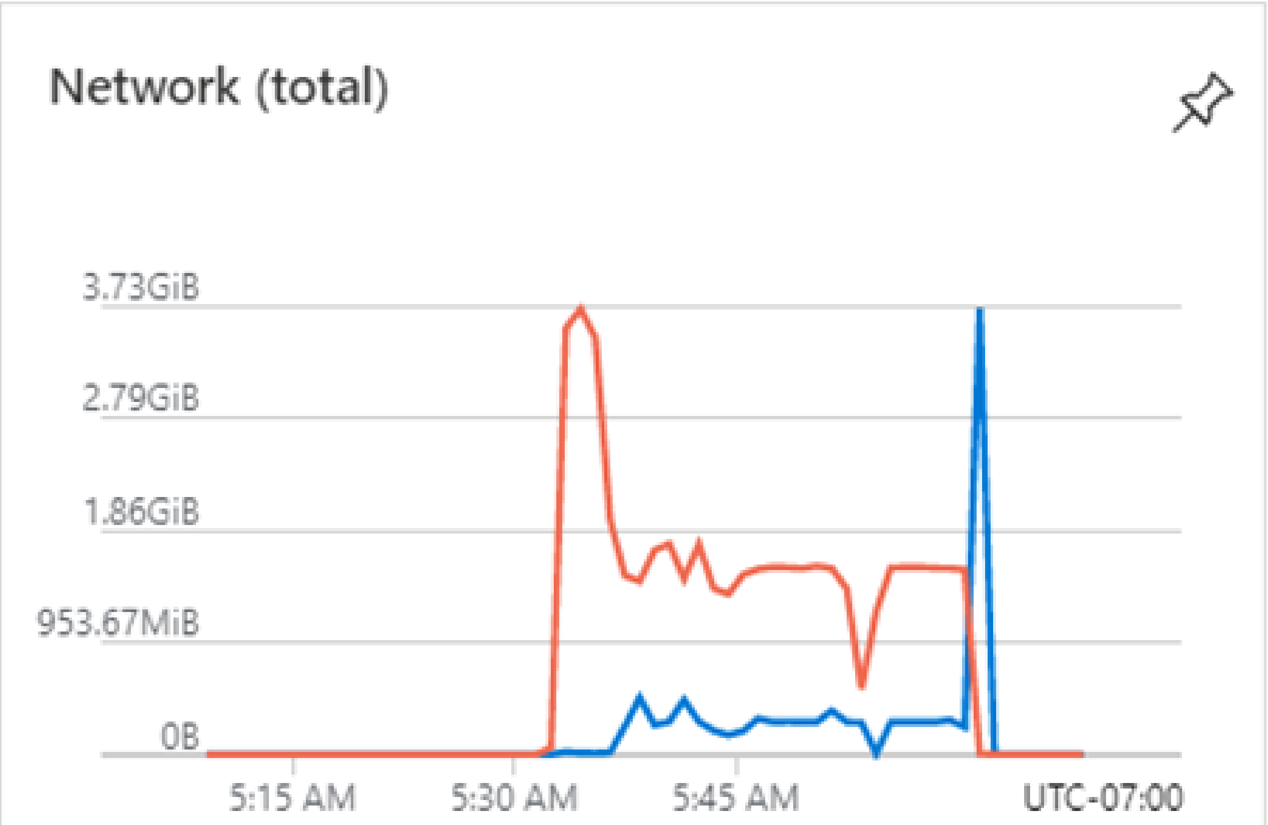
Lustre Cluster

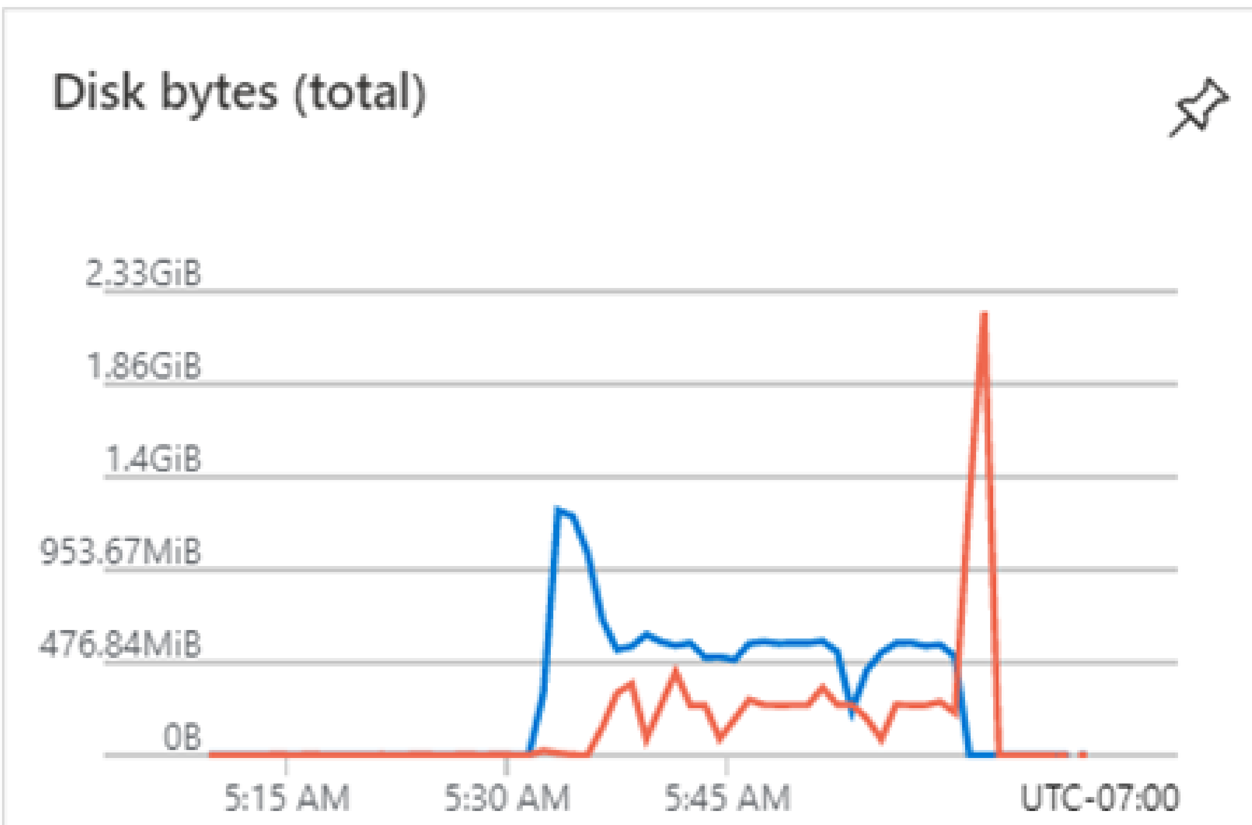
# Azure HPC File System Portfolio

| | NFS Blobs* | NFS Files* | HPC Cache | Azure NetApp Files | *aaS PVFS | Cray ClusterStor |
|---|---|---|---|---|---|---|
| File System | NFSv3 | NFSv4.1 | NFSv3<br>NFSv4.1*<br>SMB2.1* | NFSv3<br>NFSv4.1<br>SMB | Lustre<br>BeeGFS | Lustre |
| Characteristics | Large size<br>Medium throughput<br>Sequential access<br>Read-heavy | Medium size<br>Medium throughput<br>Random access<br>In-place updates | Large file<br>Sequential reads<br>Optimizes latency and throughput<br>Caches multiple source NAS environments | Medium size<br>Medium throughput<br>Random access<br>In-place updates<br>Low latency | Large size<br>High throughput<br>Sequential access<br>10-15 MiB/s per provisioned TiB | Very large size<br>High throughput<br>Sequential access<br>10-15 MiB/s per provisioned TiB |
| Use cases | Legacy NFS apps<br>Backup and archive<br>Analytics | Shared app data<br>Databases<br>Container storage<br>Home directories | HPC up to 8GB/s<br>Read heavy<br>Cloud burst from on-prem NAS<br>Multi-source file system (on-prem and in-cloud) | Enterprise application migrations<br>NFS home dirs | Built to specs<br>Durable or Ephemeral options | Long-lived HPC jobs (weeks/months)<br>Data protection and tiering in Blob storage |
| File system details | 5 PiB<br>100K IOPS (premium)<br>12.5GB/s throughput | 100TiB<br>100K IOPS<br>10GB/s | 50TiB cache<br>240k IOPS<br>8 GB/s | 100 TiB<br>300K IOPS<br>4.5GB/s | <1PB<br>TBD IOPS<br>up to 200GB/s | 5PiB, 15PiB, 45 PiB<br>240k IOPS<br>200GB/s |
| Region availability | Broad | Broad | Broad | Select | Broad | On-demand |

# CMAQ I/O Activity

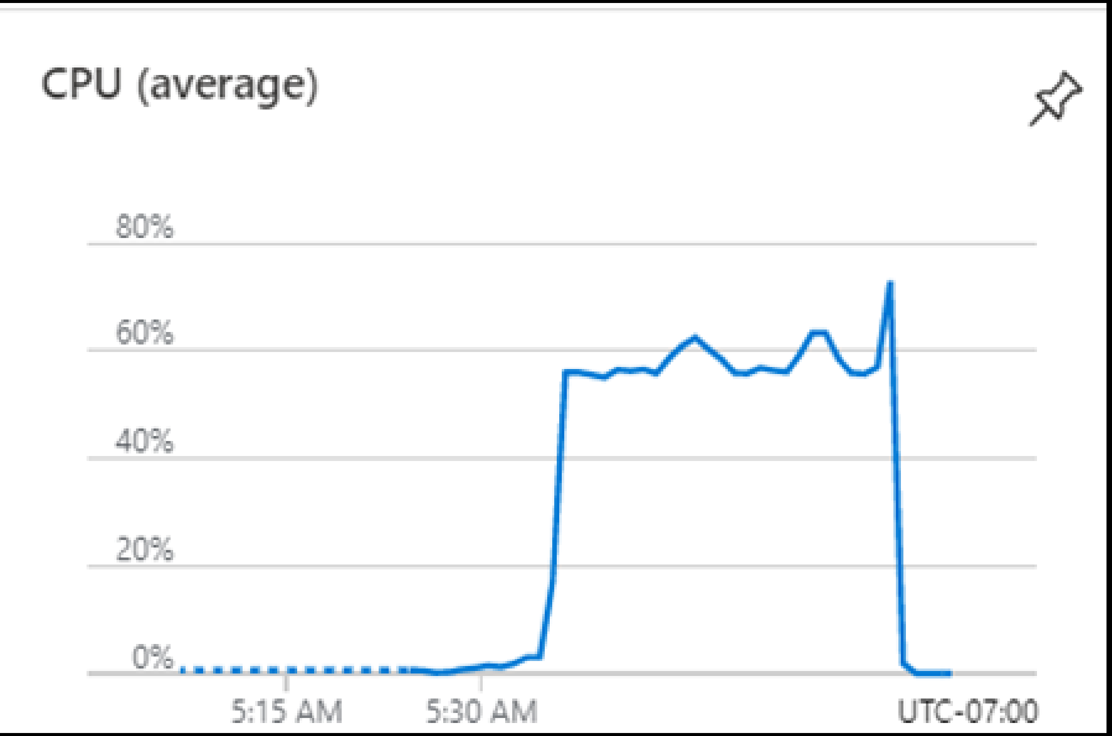
## Stand Alone NFS Server



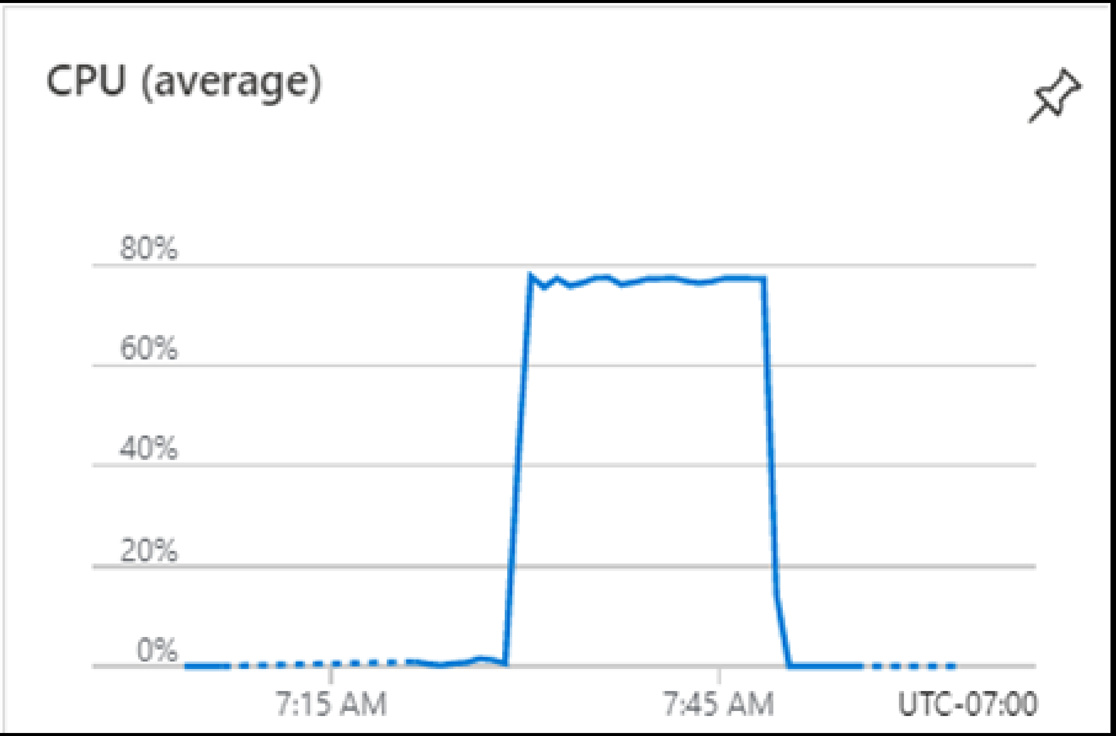



# CMAQ CPU Activity – Average across 3 Compute Nodes

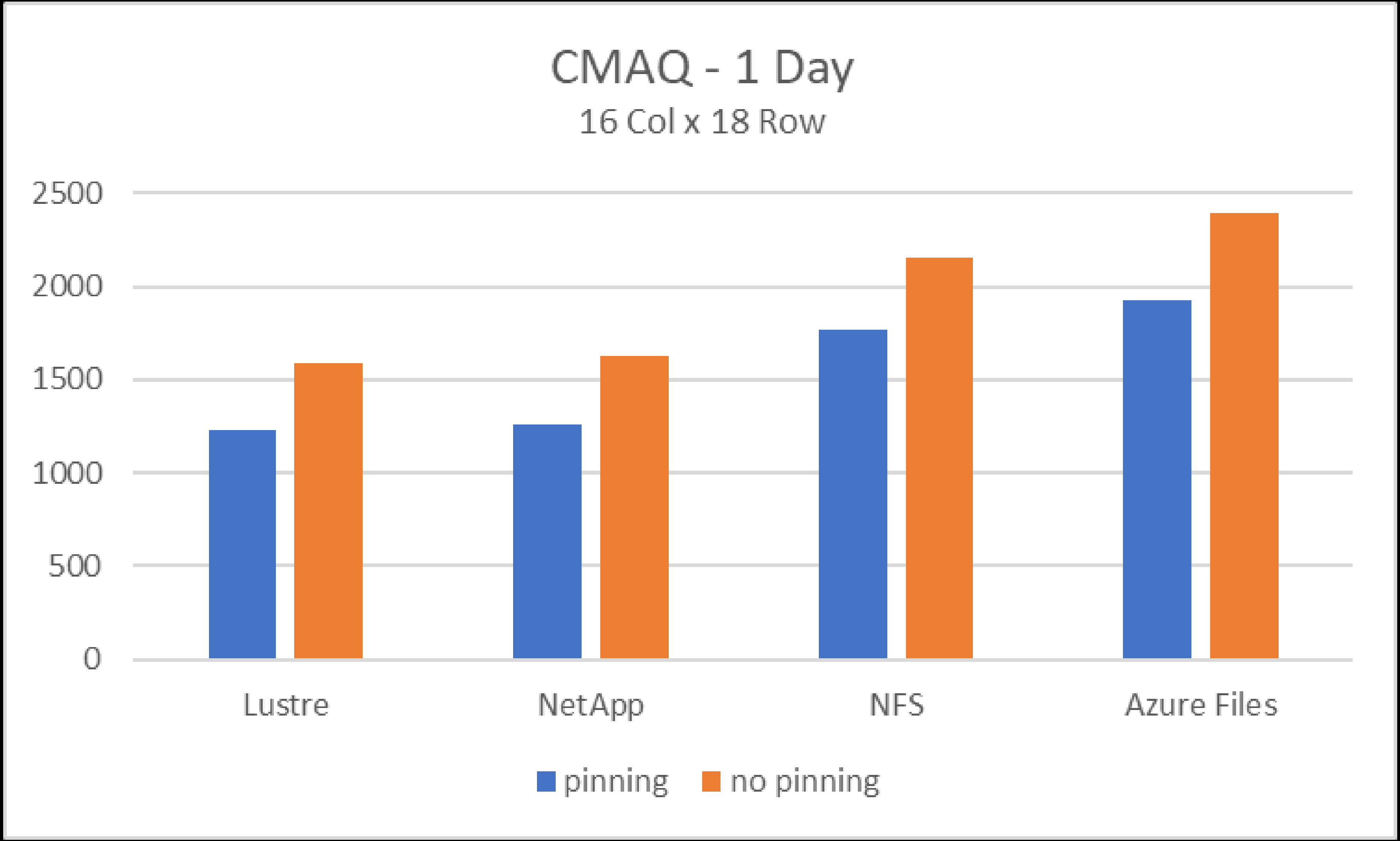
Stand Alone NFS Server




Lustre



# Test Results – 3 nodes, 96 ppn



# Questions?