Best Practices for NWP in the cloud

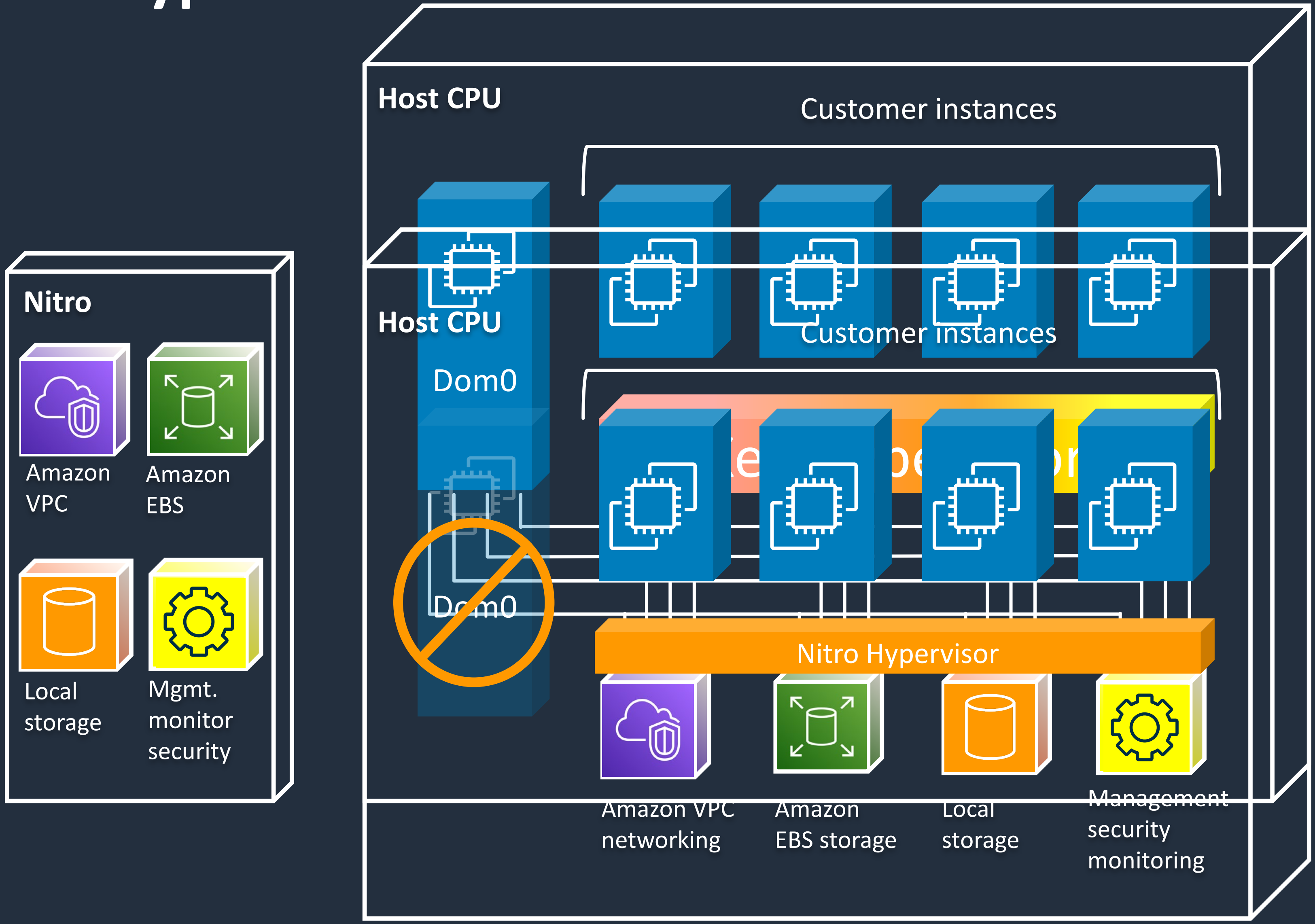
Timothy Brown
Principal Solutions Architect

Karthik Raman
Principal HPC Applications Engineer

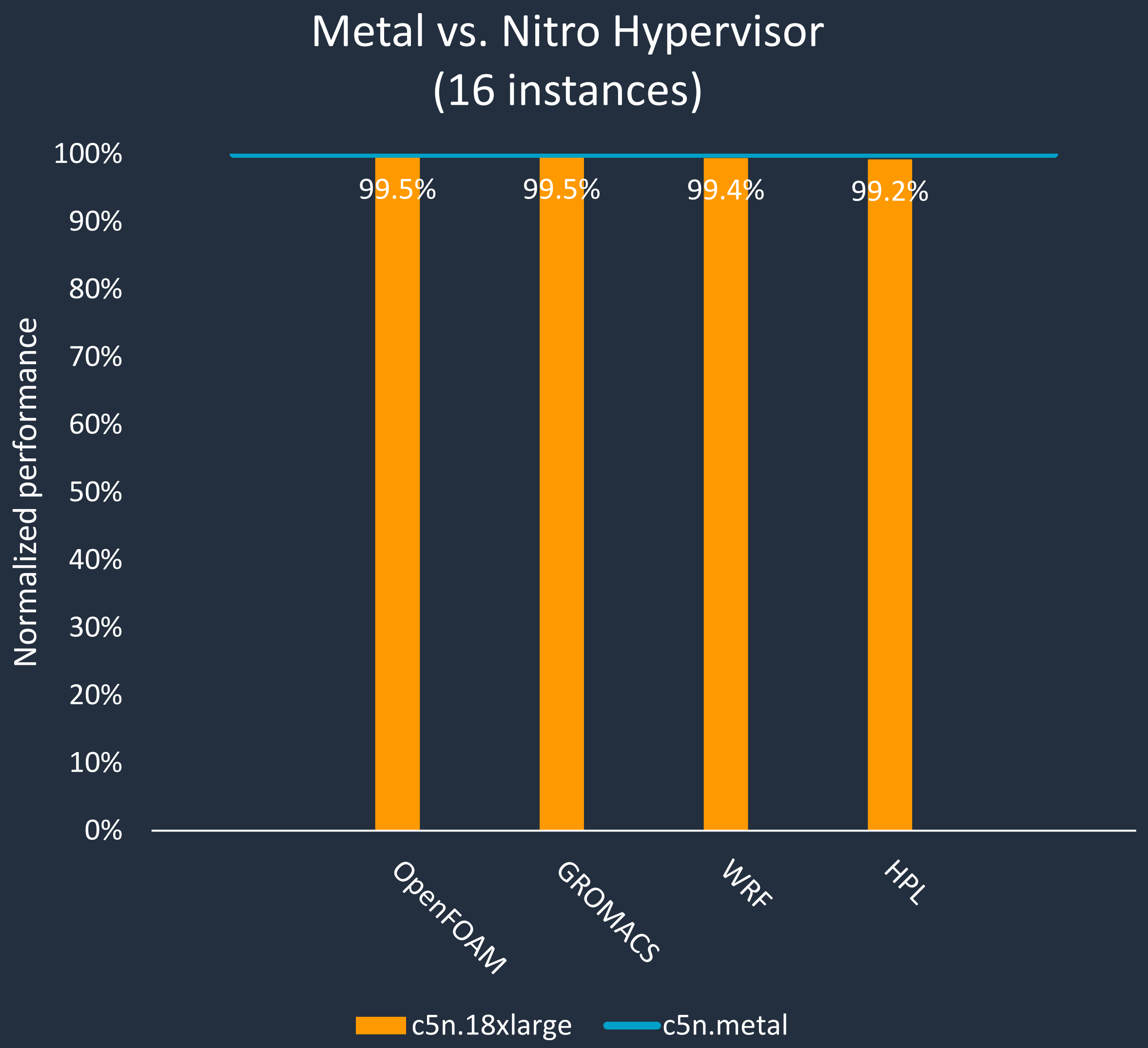
HPC Building Blocks on AWS

Performance at scale	AWS Nitro System	Amazon EC2	Elastic Fabric Adapter	Amazon FSx
Access and Orchestration	AWS Batch	AWS ParallelCluster	NICE DCV	

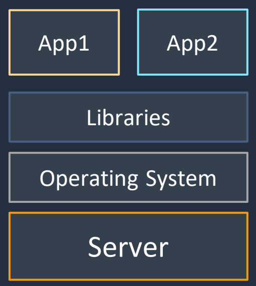
Evolution of hypervisors



The AWS Nitro System



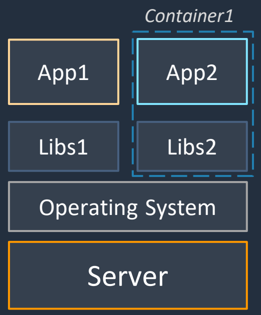
Compute: Multiple Levels of Abstraction



Traditional

Classic bare metal or VM
(Amazon EC2)

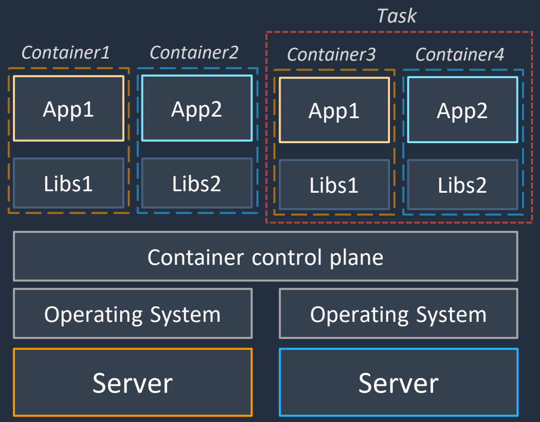
- Known environment
- Low portability



Container

Docker, Singularity, Shifter, Charliecloud…

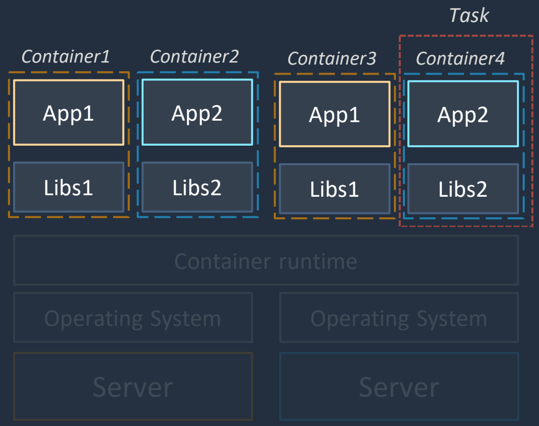
- Same env, with more portability
- Still an HPC system



Orchestrator

Abstracts infrastructure from runtime. Initially for services. Mixed serverless.
(Kubernetes, Amazon ECS/EKS, Docker Swarm)

- Can run MPI
- Containers only



Serverless

Infrastructure managed by cloud provider, jobs submitted as containers. *(AWS Lambda)*

- Consumption model
- Code and Containers,
  no infrastructure exposure

Cloud Provider Operational Responsibility

Change Compute Resources to Match Workload

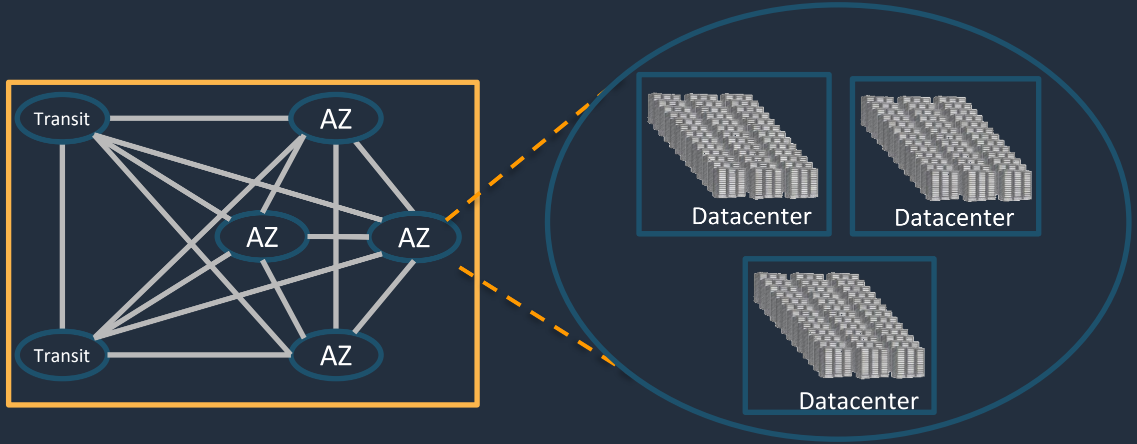
(not the other way around)

	Hpc7a.96xlarge	Hpc6id.32xlarge	Hpc7g.16xlarge
CPU	AMD (Genoa)	Intel (Ice Lake)	ARM (Graviton 3)
Cores	192 cores	64 cores	64 cores
Clock Speed*	3.6 GHz	3.5 GHz	2.6 GHz
Memory	768 GB	1024 GB	128 GB
Local Disk		4x 3800GB NVMe	
Network	300 Gb/s	200 Gb/s	200 Gb/s

* GHz figures listed are sustained all-core turbo frequencies for AMD and Intel

Regions & Availability Zones

Compute where it makes the most sense



Build for availability; understand locality

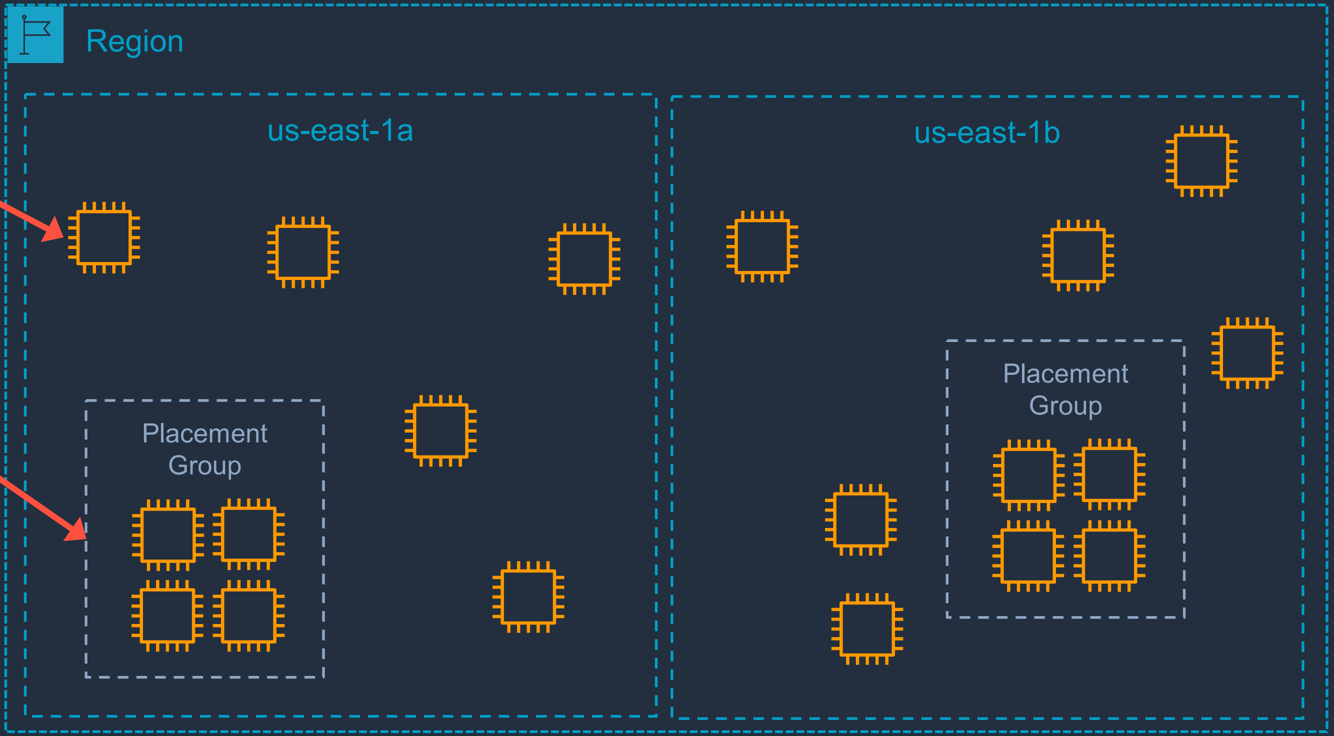
Some cloud services are region-wide others may be localized

Influencing instance placement with Placement Groups

Instances can be distributed within an Availability Zone (AZ)

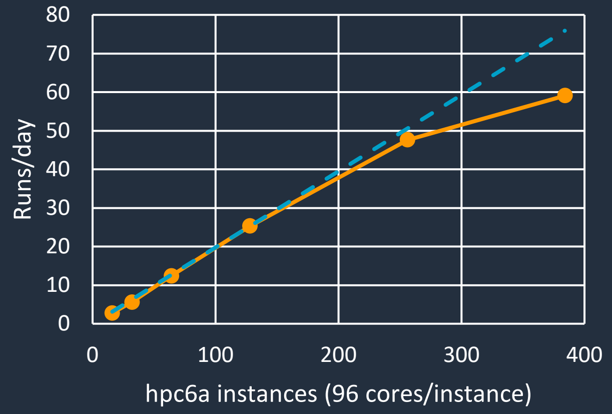
You can logically group them within an AZ using a Placement Group

Placement Group are **strongly** recommended for tightly coupled workloads (HPC, ML)



Elastic Fabric Adapter: Network Built to Scale

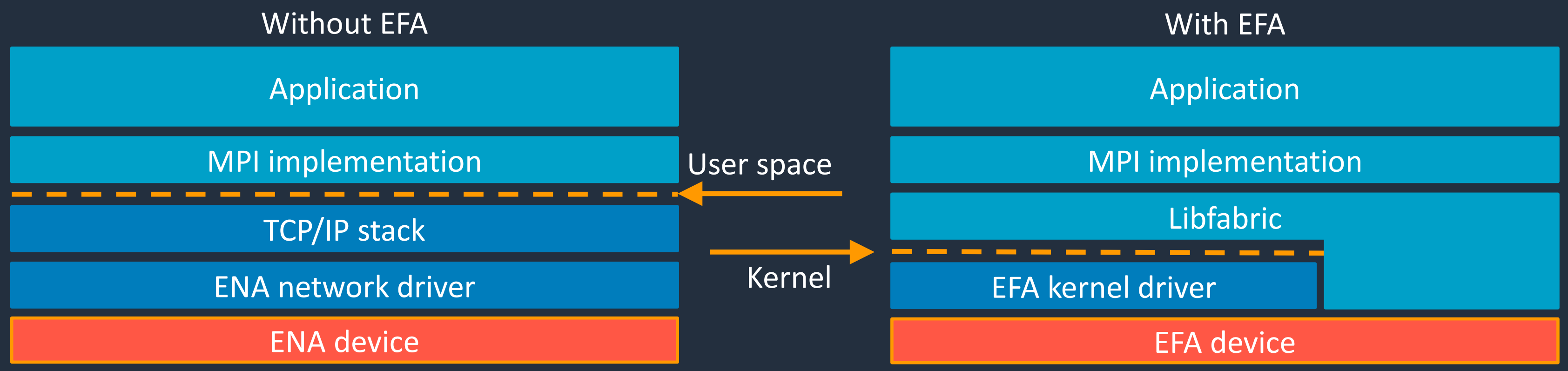
MPAS Hurricane Laura
EFA Scaling Study



Scaling for tightly-coupled workloads

- ✓ OS bypass
- ✓ GPUdirect and RDMA
- ✓ Libfabric core supports wide array of MPIs and NCCL

- High-speed/low-latency
- Cloud-scale congestion control
- Up to 3200 Gbps bandwidth
- Uses Scalable Reliable Datagram (SRD)



> “The Hpc6a, featuring AMD EPYC 3rd generation processors, combined with the EFA networking capability provides us a 60% performance improvement over alternatives, while also being more cost efficient.”
> – Dan Nord, SVP and Chief Product Officer at Maxar Technologies

Amazon FSx for Lustre

FULLY MANAGED SHARED STORAGE BUILT ON THE WORLD'S MOST POPULAR HIGH-PERFORMANCE FILE SYSTEM

FSx

Sub-ms latencies, **hundreds of GB/s of throughput,** millions of IOPS

Concurrent access for thousands of instances and **100,000s of cores**

Cost-optimized file systems with HDD and SSD storage options

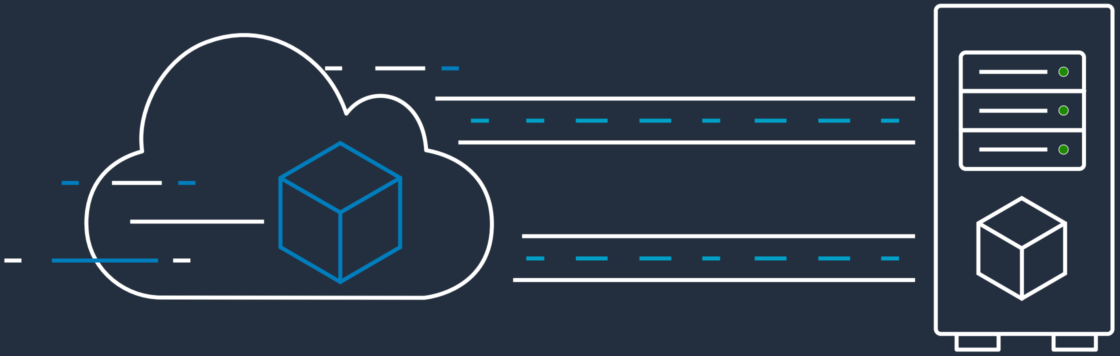
Flexible deployment options for short- and longer-term workloads

Learn more: Amazon FSx for Lustre, https://aws.amazon.com/fsx/lustre/

aws

AWS ParallelCluster

HPC Clusters and integrated services, on-demand



Integrated with AWS services you need

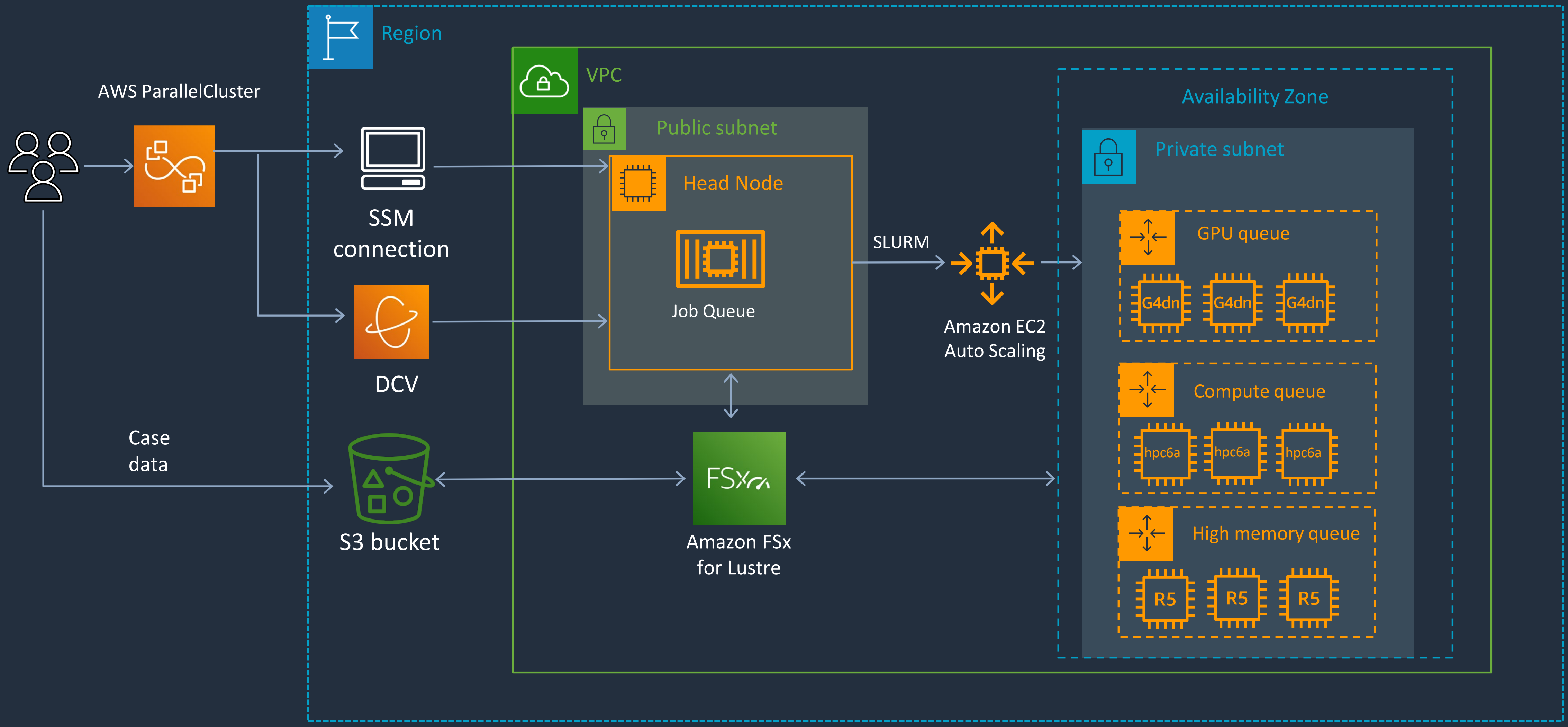
Highly-performant file systems

Amazon EC2 instances

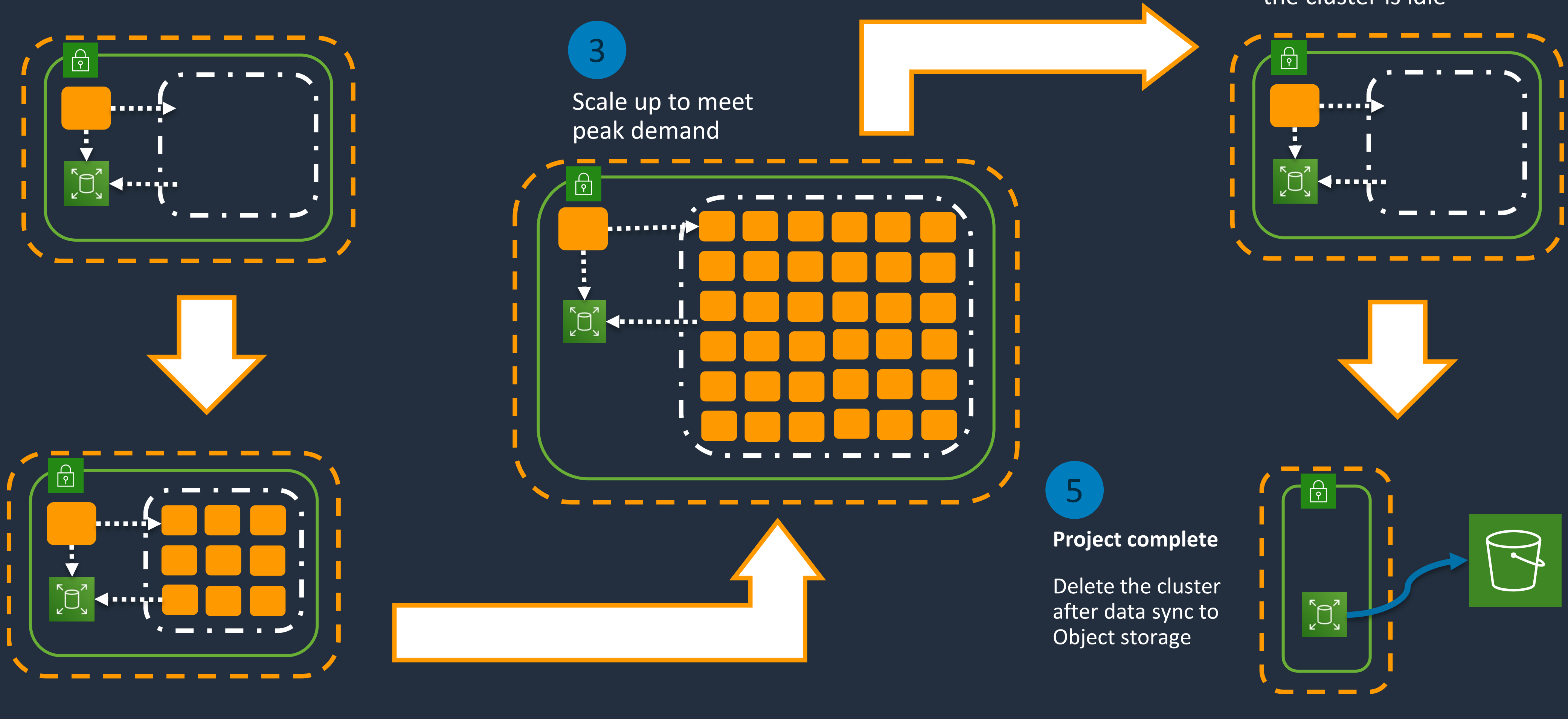
EFA

NICE DCV

AWS ParallelCluster Common Architecture



Automatic resource scaling



1

Cluster created

No Compute nodes allocated

2

Running jobs automatically allocates instances

3

Scale up to meet peak demand

4

Scale down when the cluster is idle

5

Project complete

Delete the cluster after data sync to Object storage

Worldwide Collaboration on Weather and Climate HPC

Global Weather & Climate Model Cloud Enablement

- WRF
- FV3GFS
- MPAS
- Unified Model
- Harmonie
- ICON / ICON-CLM
- GEM
- CESM
- E3SM

NCAR | National Center for
UCAR | Atmospheric Research

JCSDA

Public Sector and Commercial Deployments

MAXAR

NOAA

Australian Government
Bureau of Meteorology

DTN

Danmarks
Meteorologiske
Institut

Research and Open Data Pipelines

NOAA

NASA

aws

BoM Testcase, Priorities, and Goals

Unified Model Testcase Details

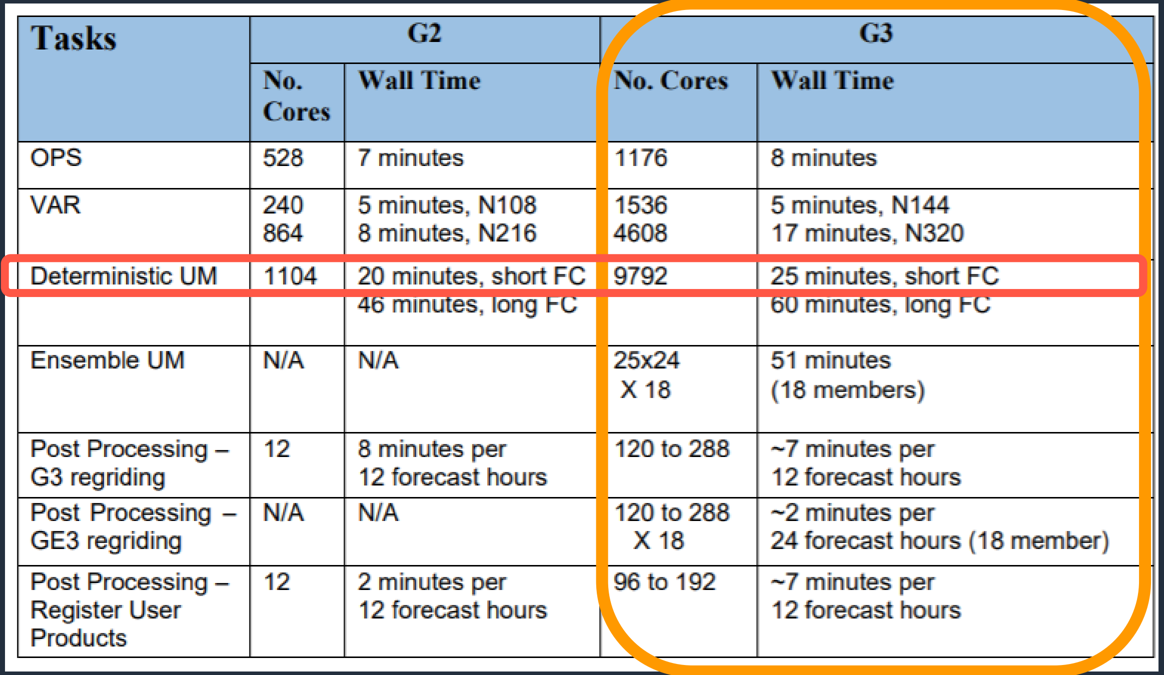
- APS3, N1024L70
- Forecast length- 72hrs (3 days)
- APS3 Grid Points- 1536 latitude x 2048 longitude
- APS3 Grid Spacing- 12km

Priorities

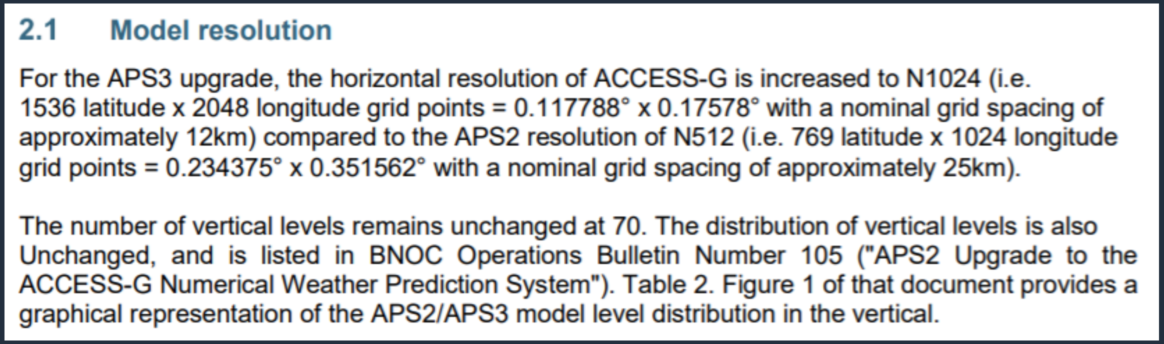
- Compare Amazon EC2 instance price/performance
- Containerize UM NWP runs using Singularity
- Optimize decomposition parameters

Goals

- 3 day forecast in < 18mins (compute + file I/O)
- This requirement is derived from BoM operational 3.5 day forecast taking < 22 mins (avg), 25 mins (wc)
- Identify options for lowest cost to results while meeting performance requirement



Tasks	G2		G3	
	No. Cores	Wall Time	No. Cores	Wall Time
OPS	528	7 minutes	1176	8 minutes
VAR	240 864	5 minutes, N108 8 minutes, N216	1536 4608	5 minutes, N144 17 minutes, N320
Deterministic UM	1104	20 minutes, short FC 46 minutes, long FC	9792	25 minutes, short FC 60 minutes, long FC
Ensemble UM	N/A	N/A	25x24 X 18	51 minutes (18 members)
Post Processing – G3 regriding	12	8 minutes per 12 forecast hours	120 to 288	~7 minutes per 12 forecast hours
Post Processing – GE3 regriding	N/A	N/A	120 to 288 X 18	~2 minutes per 24 forecast hours (18 member)
Post Processing – Register User Products	12	2 minutes per 12 forecast hours	96 to 192	~7 minutes per 12 forecast hours



2.1 Model resolution

For the APS3 upgrade, the horizontal resolution of ACCESS-G is increased to N1024 (i.e. 1536 latitude x 2048 longitude grid points = 0.117788° x 0.17578° with a nominal grid spacing of approximately 12km) compared to the APS2 resolution of N512 (i.e. 769 latitude x 1024 longitude grid points = 0.234375° x 0.351562° with a nominal grid spacing of approximately 25km).

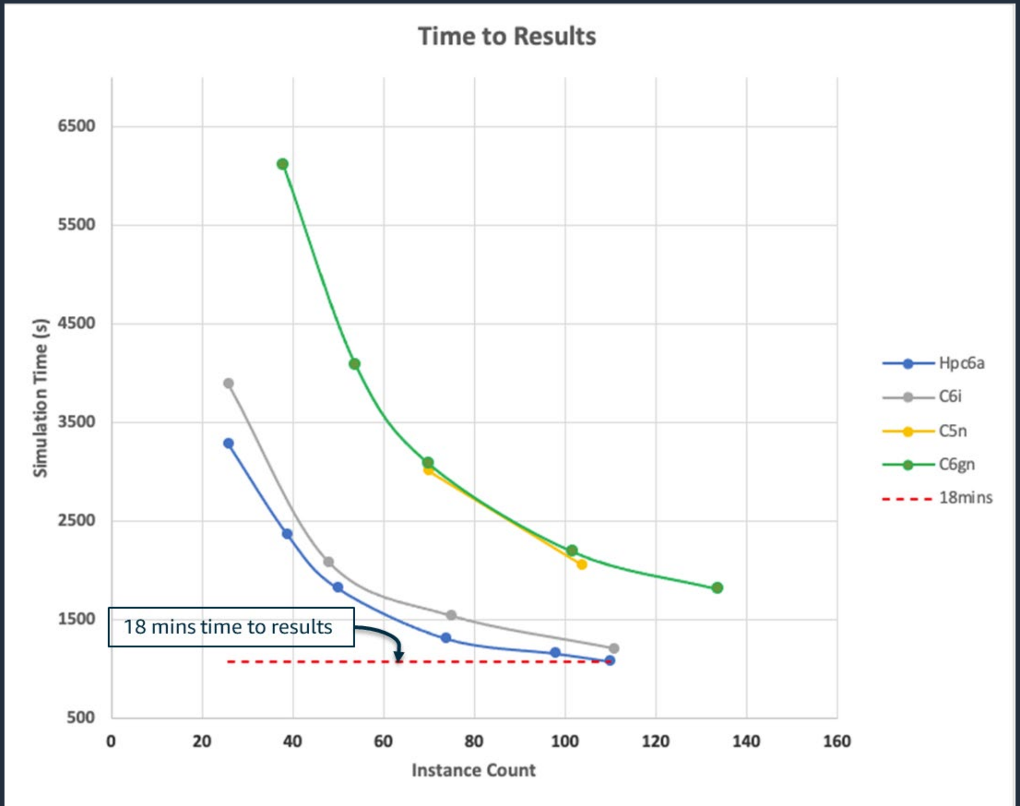
The number of vertical levels remains unchanged at 70. The distribution of vertical levels is also Unchanged, and is listed in BNOC Operations Bulletin Number 105 ("APS2 Upgrade to the ACCESS-G Numerical Weather Prediction System"). Table 2. Figure 1 of that document provides a graphical representation of the APS2/APS3 model level distribution in the vertical.

Source- http://www.bom.gov.au/australia/charts/bulletins/opsbull_G3GE3_external_v3.pdf

Australia BoM: Up to 78% better price performance

Australian Government
Bureau of Meteorology

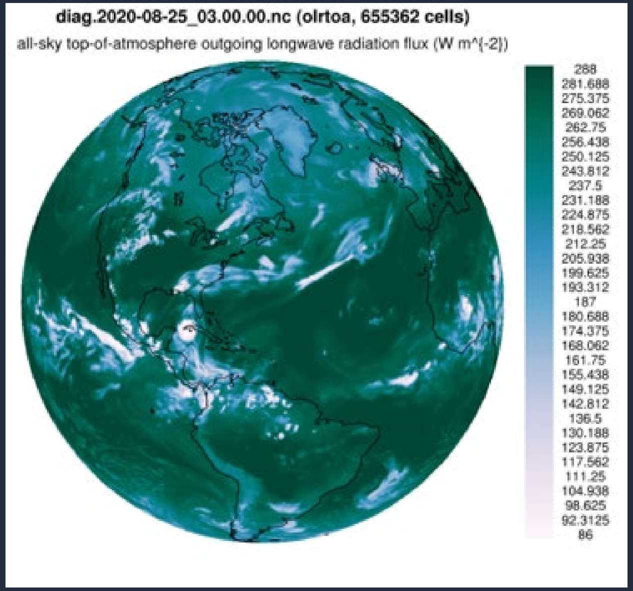
- Amazon EC2 Hpc6a instance shown to be a viable choice for BoM's NWP use cases based on results from the G3 (APS3, N1024L70) 72hr testcase
- Hpc6a achieves the **18min time to results requirement with ~110 instances (10,560 cores)**. Additionally, Hpc6a achieves **up to 59% lower cost and 78% better price/performance than comparable C-family instances** (such as C6i, C5n)
- AWS's Unified Model runs on Singularity show that there is less than 1% performance variation between containerized and non-containerized options.



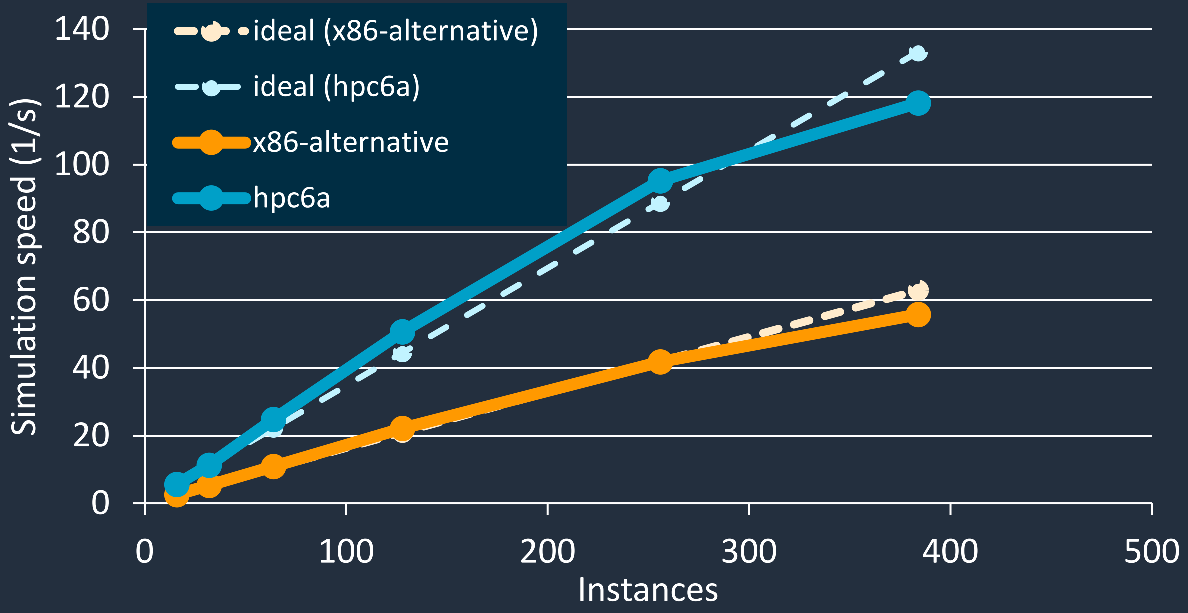
DTN: Enabling High-resolution Weather Modeling

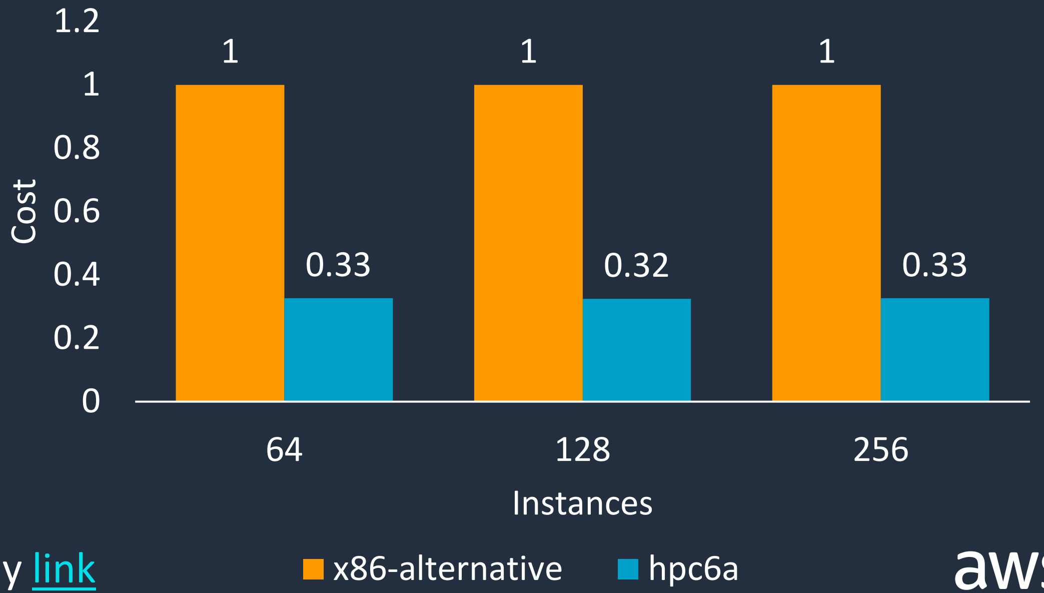
“Our collaboration with AWS allows us to better serve our customers with high-resolution weather prediction systems that feed analytics engines. ***We’re very excited to see the price/performance of Hpc6a and we expect this to be our go-to Amazon EC2 instance choice for HPC workloads going forward.”***

- Lars Ewe, Chief Technology Officer, DTN


diag.2020-08-25_03.00.00.nc (olrtoa, 655362 cells)

all-sky top-of-atmosphere outgoing longwave radiation flux (W m^{-2})

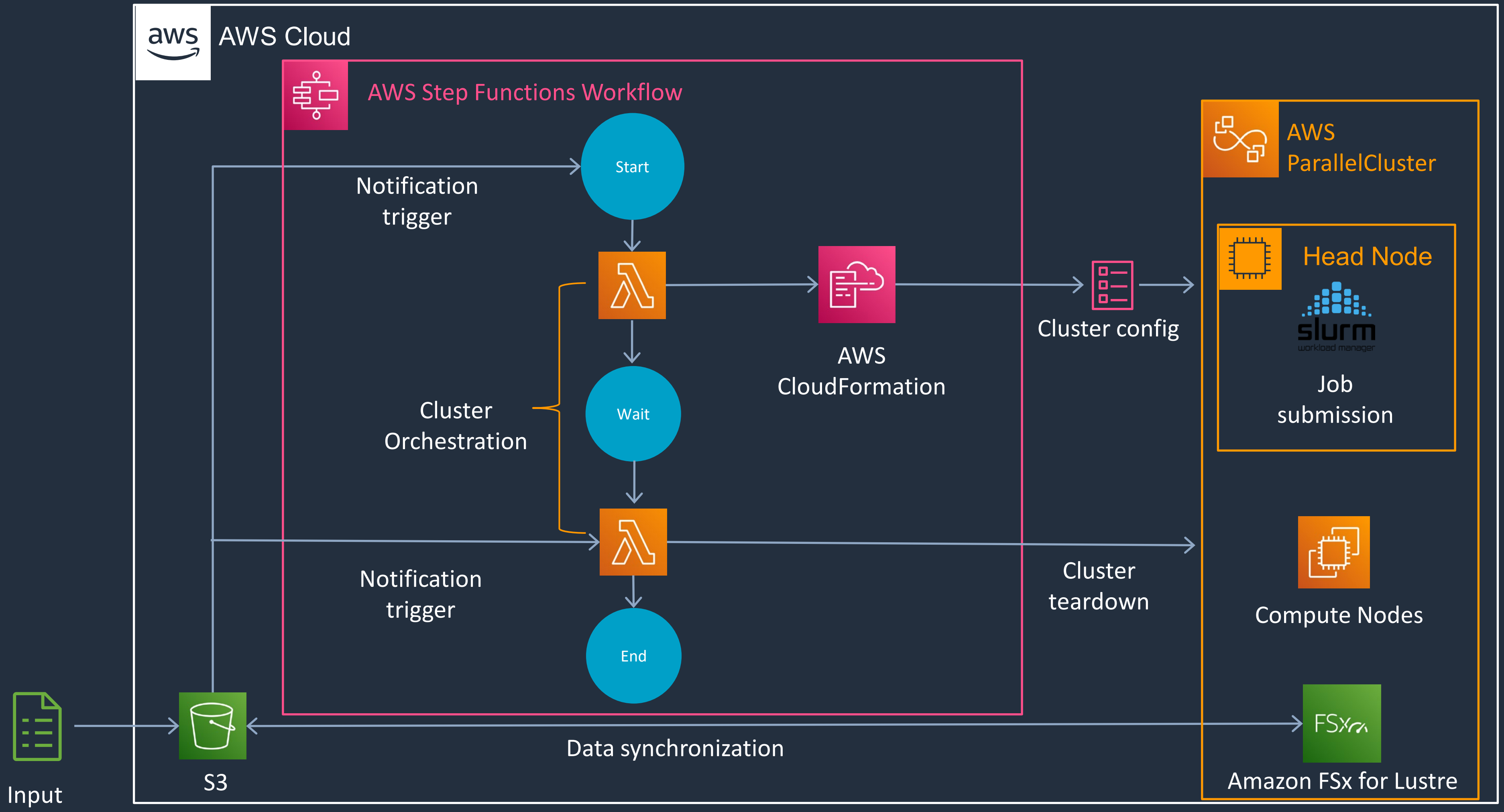




2022 DTN AWS Case Study link

Event Driven

https://github.com/aws-samples/event-driven-weather-forecasts



Resources

- NWP workshop: https://catalog.workshops.aws/nwp-on-aws/
- CMAQ workshop: https://catalog.workshops.aws/cmaq-tutorial
- AWS Batch: https://batch.hpcworkshops.com/
- SC23 Tutorial Monday 13th Nov:
  https://sc23.supercomputing.org/presentation/?id=tut144&sess=sess238

Questions?