



Suitability of Commercial Clouds for NASA's HPC Applications

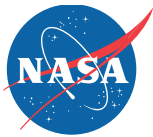
Robert Hood, InuTeq

NASA Advanced Supercomputing Division

NASA Ames Research Center

HPC User Forum

The HECC Project at NASA Ames



High End Computing Capability

- Facilitating Science/Engineering at NASA
- 4 compute systems, including:
 - Pleiades: #24 in Top500; #14 in HPCG
 - Electra: #43 & #37 (both should go up)
- 1300+ users from across Agency Directorates
 - Aeronautics, Human Exploration, Science
- ~900 MPI applications (+ non-MPI codes)

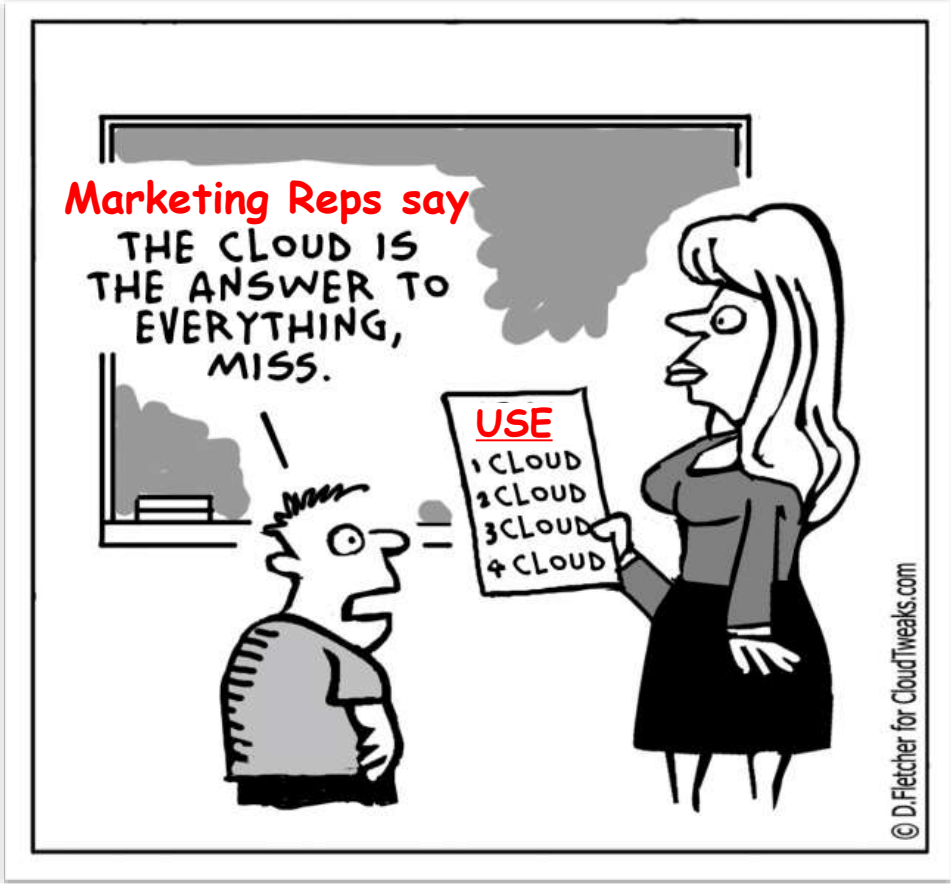


Goal: Maximize resource delivery to users

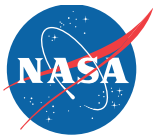
- Strive for >80% utilization of resources
- Embrace new facility technologies
- Continually evaluate procurement options



NASA HQ: Is HECC doing the best it can?



Answering the Question: A Trade Study



Key Questions:

- Would clouds be a cost-effective replacement for on-premises resources?
- Under what conditions could cloud usage be cost effective?

Approach: Compare costs of running NASA workload on premises & in cloud

Evaluation Systems

- Existing resources on premises at NASA InfiniBand interconnect
- Amazon Web Services (AWS) Ethernet interconnect
- Penguin-On-Demand (POD) InfiniBand, OmniPath interconnects

Benchmarks

- NAS Parallel Benchmarks (NPBs): Classes C and D
- 6 full-sized applications:
 - *Weather/Climate*: FVCore, ECCO (really MITgcm), WRF/nuWRF
 - *Science*: Athena++, Enzo
 - *CFD*: OpenFoam *representing NASA applications that are export controlled and couldn't be used for security reasons*

Trade Study Approach, continued



Cost Basis:

- Use worst-case cost estimate for on premises: the full cost
 - Compute cost of delivering a System Billing Unit (SBU) of work:
$$\text{CostPerSBU} = (\text{Total HECC Budget}) \div (\text{number of SBUs delivered to Agency})$$
 - Cost of run is: *(number of SBUs used in run) × CostPerSBU*
 - Note: includes all hardware, software, power, maintenance, staff, and facility costs
- Use best-case cost estimate for cloud: the compute-only cost
 - AWS: use estimate of best spot price (30% of on-demand price)
 - POD: use published rates (discounts likely available, esp. for volume)
 - Note: does not include costs for storage, software licenses, bandwidth, HECC staff, etc.

Results and Analysis

- Do a performance and cost analysis for jobs running on similar processor types
- Produce findings and identify actions for HECC

Comparison of Compute Resources Used



	HECC	AWS (Costs are Compute-Only [†])					POD
Model/Cores	Full Cost	Instance Name	On-Demand	Spot Price	GovCloud On-Demand	GovCloud Pre-Leasing	Compute Only
Haswell/18		m4.16xlarge	\$1.591	\$0.477	\$1.915	\$1.341	
Haswell/20							\$1.800
Haswell/24	\$0.534			\$0.636*			\$2.160*
Broadwell/28	\$0.646			\$0.840*			\$2.800
Broadwell/32		c4.8xlarge	\$3.200	\$0.960	\$4.032	\$2.822	
Skylake/36		c5.18xlarge	\$3.060	\$0.918**	N/A	N/A	
Skylake/40	\$1.018			\$1.020* **			\$4.400

**AWS spot prices in blue are approximations for comparison purposes and are pro-rated based on relative core counts versus HECC. The POD 24-core Haswell price has been similarly converted for comparison to HECC.*

***This estimated AWS spot price for Skylake is unlikely to be available—see Appendix I.*

†Full cost information unavailable, but expected to be significantly higher than compute-only cost.

Finding: The per-hour full cost of HECC resources is cheaper than the (compute-only) spot price of similar resources at AWS and significantly cheaper than the (compute-only) price of similar resources at POD.

NAS Parallel Benchmark Results

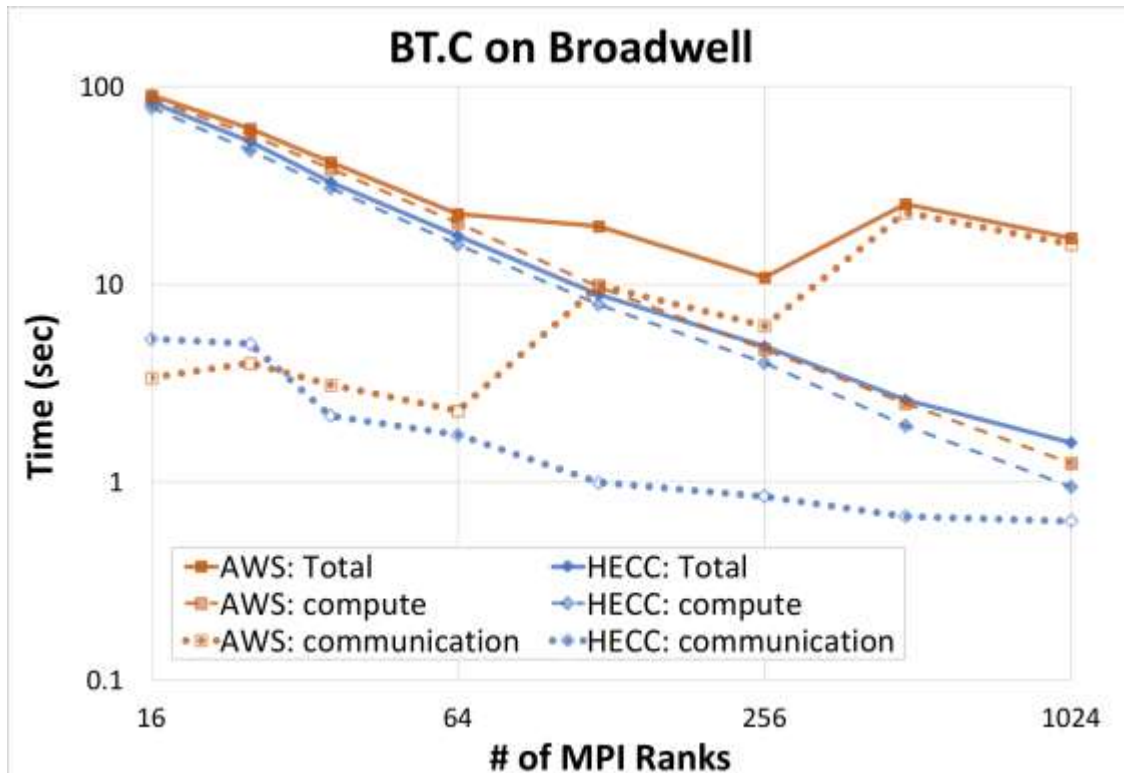


Performed runs from to 16–1024
MPI ranks to evaluate scaling

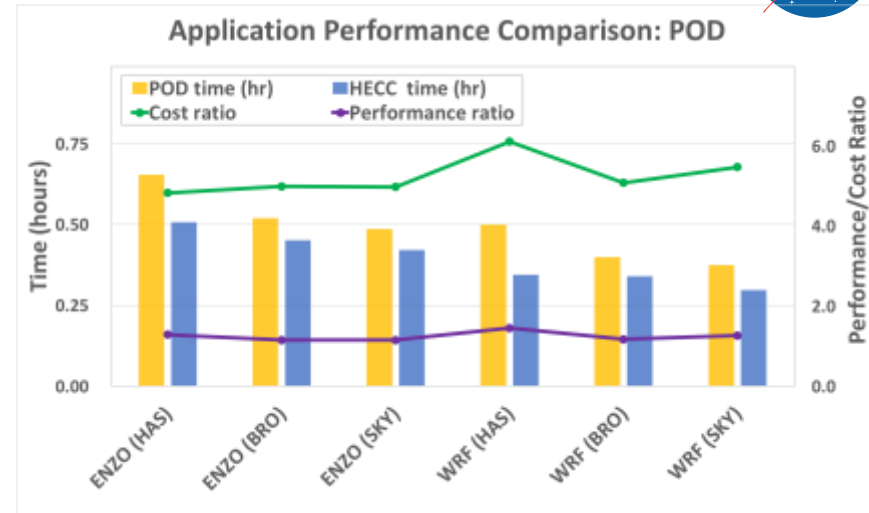
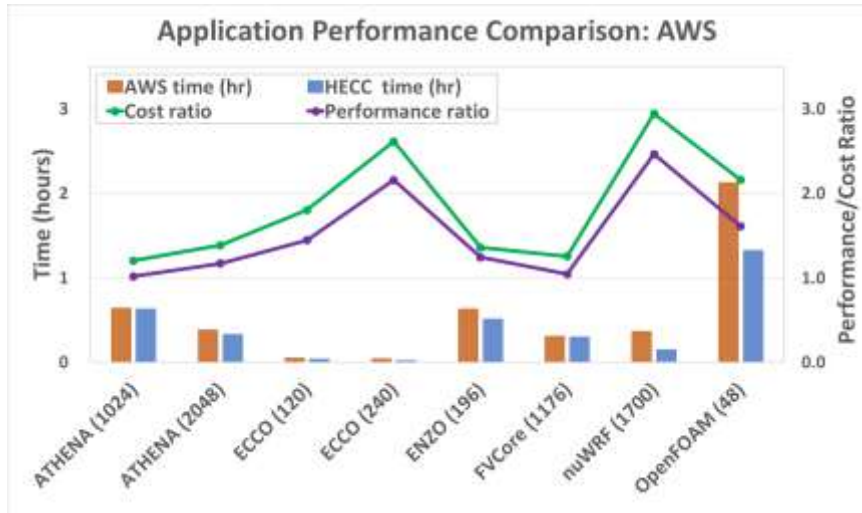
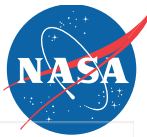
- POD performance was similar to HECC
- AWS performance was worse than HECC on communication-intensive benchmarks

Cost to run the full set of the NPBs

- AWS was 5.8–12 times more expensive than HECC, depending on the processor type used
- POD was 4.7 times more expensive than equivalent runs at HECC



Full-Sized Application Results



- Large computations took substantially more time on AWS than on HECC
- When cores/node are equal (BRO), POD performance is close to HECC's, but still lags
- AWS was 1.9x more expensive than HECC for the application workload, even with spot pricing
- POD was 5.3x more expensive than HECC for the application workload

Finding: Tightly-coupled, multi-node applications from the NASA workload take somewhat more time when run on cloud-based nodes connected with HPC-level interconnects; they take significantly more time when run on cloud-based nodes that use conventional, Ethernet-based interconnects.

Main Finding



Finding: Commercial clouds do not offer a viable, cost-effective approach for replacing in-house HPC resources at NASA

Why might others get different results?

- Our applications have different computation & communication patterns from theirs?
- Low HECC Costs
 - Equipment, power, cooling
 - Staff: size of system means lower staff costs per SBU than smaller centers
- High HECC Utilization
 - Evolution of system:
 - *Effectively same system for 10 years*
 - *User familiarity reduces ramp up costs to reach high utilization of new resources*
 - Minimize interruptions for maintenance
 - *Rolling upgrades, suspend/resume, ...*

So, are there cases in HECC where clouds might be cost effective?

Cost-Effective Uses of Cloud for NASA



When utilization would be low

- New resource types, e.g. GPU-accelerated nodes
 - Use cloud-based resources until demand rises to the point where it is more cost effective to acquire resources and move work on premises

When there are other costs to consider

- Opportunity costs associated with high utilization, e.g. longer waits in the queue
 - What if we move 1-node jobs to the cloud? Would that speed up scheduling of remaining jobs?

Real-time requirements

- e.g web services

Finding: Commercial clouds provide a variety of resources not available at HECC. Certain use cases may be cost-effective to run there.

Follow-up Work Identified



Get a better understanding of potential benefits and costs of having a portion of the HECC workload in the cloud

- Impact of 1-node jobs on scheduling of rest of workload
- Understand performance characteristics of jobs that might run there

Define a comprehensive model that allows accurate comparisons of cost of running jobs depending on resources used

Prepare for a broadening of services provided by HECC to include a portion of its workload running on commercial cloud resources

- For HECC users
- For non-HECC users

Running 1-Node Jobs in the Cloud



Motivation

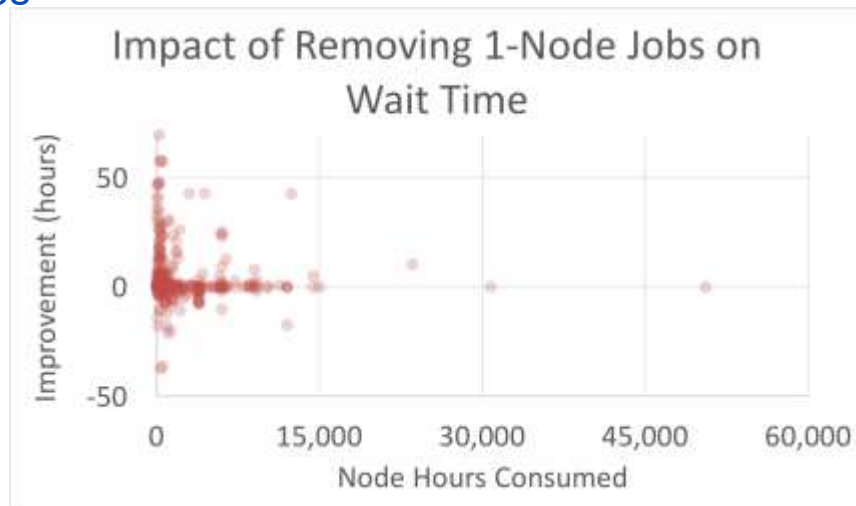
- One-node jobs not as likely to see performance impact due to commodity interconnect
- Does moving them to the cloud speed up scheduling of wider jobs?

Experiment

- Simulate scheduling of actual job mix with 1-node jobs removed
- Determine impact to starting time of remaining jobs

Results

- While some jobs started much sooner, they tended to be using relatively few nodes
- Weighted average of improvement was ~2 hours
- Not clear that benefit would be worth added cost of using cloud resources



Integrating Cloud Usage into HECC



Pilot project to move jobs from HECC resources to AWS

- HECC user logs into cloud front end to build executable
- User annotates batch scripts, indicating files that need to be staged to/from cloud
- User submits batch jobs to “cloud” queue
- PBS server moves jobs to server running at AWS; stages input files
- Server in cloud allocates resources, runs job, stages output files back
- Accounting is done manually; HECC pays
- Limited to non-export controlled codes and data (i.e. “low” security plan)

User-defined software stacks

- Security concerns with Singularity (runs as root)
- Charliecloud-based containers are unprivileged
- Deployed Charliecloud on HECC; tested on AWS as well
 - Don't yet have a solution for MPI programs running on multiple nodes

Future Work



Extend Pilot Project to Include:

- Full accounting, with account limits and automated tracking of consumption
- Moderate security plan

Extend Services to Include non-HECC Users

- They bring their own funding
- Provide user interface for defining and running jobs, moving data, etc.
- HECC would add cost-recovery fee to make this self sustaining

Cost Methodology

- Need to be able to do meaningful cost comparisons between on-premises and in-cloud in order to determine which would be most cost effective
 - Include opportunity costs, too
 - Benchmark suite
- Adjust processes for acquisition and phase out of resources to include commercial cloud

Acknowledgements



The Team:

S. Chang, R. Hood, H. Jin, S. Heistand, J. Chang, S. Cheung, J. Djomehri, G. Jost, D. Kokron

Helpful Advice from:

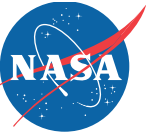
R. Biswas, B. Ciotti, L. Hogle, T. Lee, P. Mehrotra, and W. Thigpen

Resources:

- The NASA High-End Computing Capability (HECC)
- Penguin Computing made their cloud-based resources available to our evaluation at no cost and provided early access to their Skylake-based nodes

The Report: NAS Technical Report NAS-2018-001

https://www.nas.nasa.gov/assets/pdf/papers/NAS_Technical_Report_NAS-2018-01.pdf

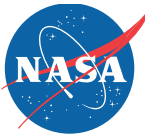


?

Backup



Full-Sized Application Results: AWS vs HECC



Benchmark	Case	NCPUS	# of HECC Haswell nodes	# of AWS c4.8xlarge (Haswell) instances	HECC time (sec)	HECC full cost	AWS time (sec)	AWS Oregon compute cost	AWS Gov compute cost
ATHENA++	SBU2	1024	43	57	2268	\$14.48	2298	\$57.89	\$69.68
ATHENA++	SBU2	2048	86	114	1177	\$15.03	1374	\$69.22	\$83.32
ECCO	NTR1	120	5	7	120	\$0.09	173	\$0.54	\$0.64
ECCO	NTR1	240	10	14	65	\$0.10	140	\$0.87	\$1.04
ENZO	SBU2	196	9	11	1827	\$2.44	2266	\$11.02	\$13.26
FVCore	SBU1	1176	49	66	1061	\$7.72	1104	\$32.20	\$38.76
nuWRF	SBU2	1700	71	95	529	\$5.58	1302	\$54.66	\$65.80
OpenFOAM	Channel395	48	2	3	4759	\$1.41	7646	\$10.14	\$12.20
OpenFOAM	Channel395	144	6	8	12547	\$11.17	20771	\$73.44	\$88.39
OpenFOAM	Channel395	288	12	16	10194	\$18.16	23013	\$162.73	\$195.87
Total Cost						\$76.18		\$472.71	\$568.96
Estimated AWS spot cost (30% of on-demand cost)								\$141.77	
Estimated AWS pre-leasing cost (70% of US-gov cost)									\$398.27

- Large computations took substantially more time on AWS than on HECC
- AWS was 1.9x more expensive than HECC for the application workload
 - Even with optimistic assumptions about getting spot pricing

Full-Sized Application Results: POD vs HECC



APP	Case	NCPUS	# of HECC nodes	# of POD nodes	HECC time (sec)	Total HECC full cost	POD time (sec)	Total POD compute cost
ENZO (HAS)	SBU2	196	9	10	1827	\$2.44	2355	\$11.78
ENZO (BRO)	SBU2	196	7	7	1625	\$2.04	1870	\$10.18
ENZO (SKY)	SBU2	196	5	5	1519	\$2.15	1751	\$10.70
WRF (HAS)	SBU1	384	16	20	1243	\$2.95	1802	\$18.02
WRF (BRO)	SBU1	384	14	14	1225	\$3.08	1436	\$15.64
WRF (SKY)	SBU1	384	10	10	1069	\$3.02	1352	\$16.52
Total Cost						\$15.68		\$82.84

- When cores/node are equal, POD performance is just a little bit worse than HECC's
- POD was 5.3 times more expensive than HECC for the application workload

Finding: Tightly-coupled, multi-node applications from the NASA workload take somewhat more time when run on cloud-based nodes connected with HPC-level interconnects; they take significantly more time when run on cloud-based nodes that use conventional, Ethernet-based interconnects.