COMMUNITY CLUSTER PROGRAM TOWN HALL- SUMMER 2020

Preston Smith – Executive Director, Research Computing

PURDUE UNIVERSITY | Information Technology

Today's Agenda

- Value Proposition of institution's investment in research computing
- Program Statistics
- Research Data
- Controlled Computing
- Cluster Lifecycles
- Value-add features on clusters
- Datacenter News
- FY21 Plans
- **2020 Cluster - Bell**
- Future Outlooks
- Discussion Topics

Value Proposition

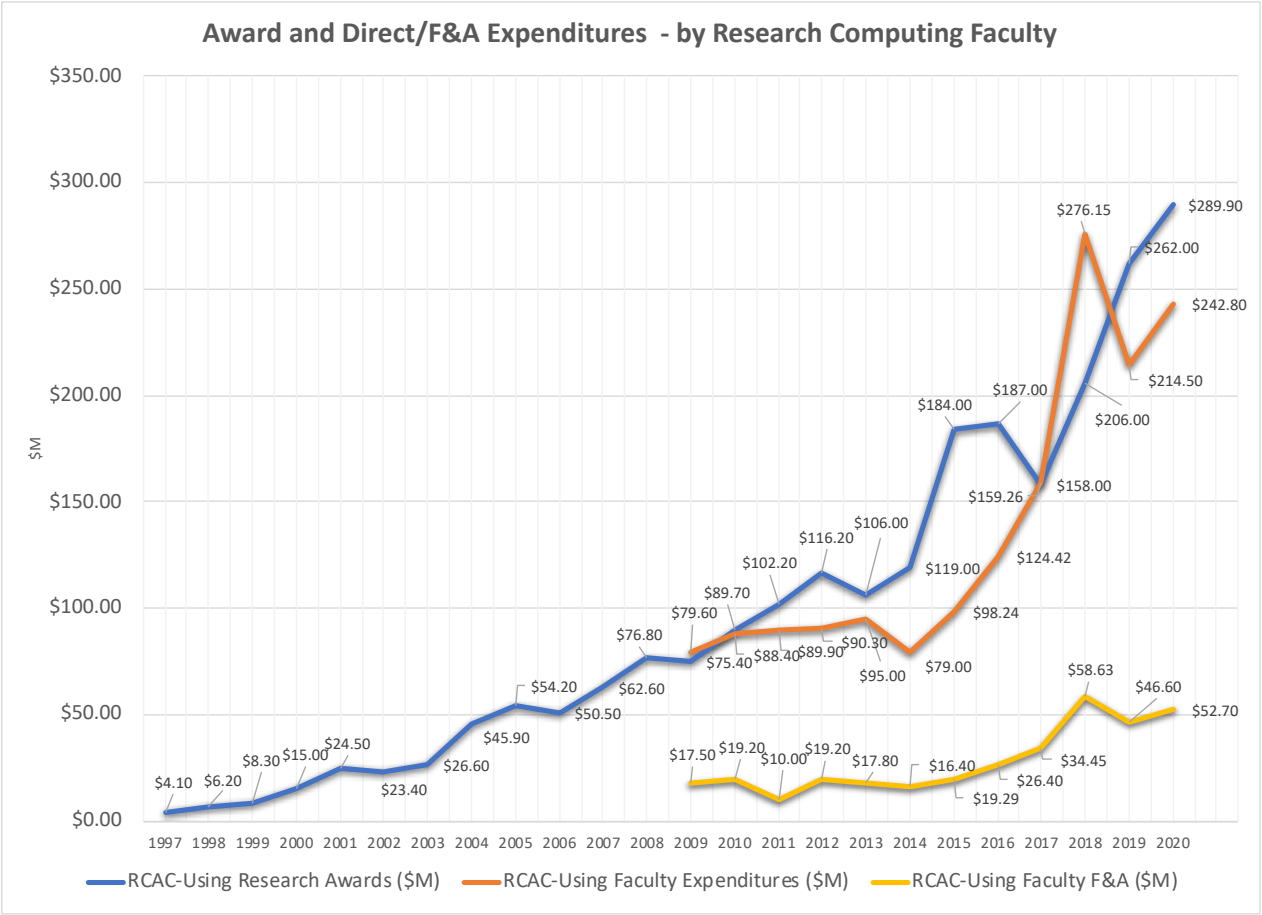
Last Year's Governance Meeting, Summarized

- Charge given to RCAC leadership and faculty governance group to articulate the value of why Purdue has to invest in these facilities
- A small group worked for several months to define metrics, analyze data, and craft the message
- Socialized among executive leadership

Annual Report:
https://www.rcac.purdue.edu/about/impact/2019_impacts.pdf

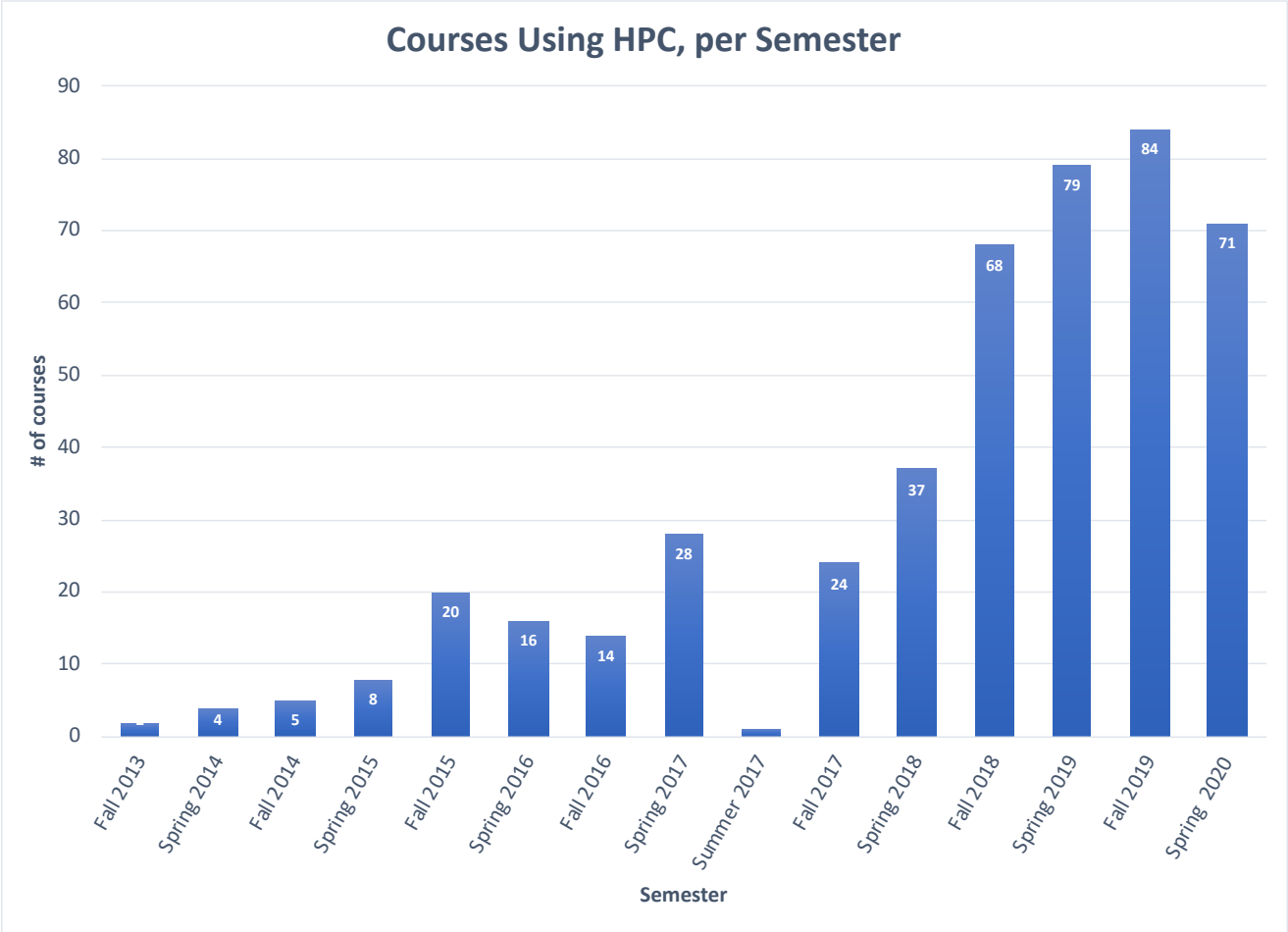
High Performance Computing is required by Purdue researchers

RCAC users were responsible for 55% of all 2020 research expenditures at Purdue ($242M), and won 56% of all awards ($289M)



High Performance Computing is a crucial capability for STEM education and data sciences.

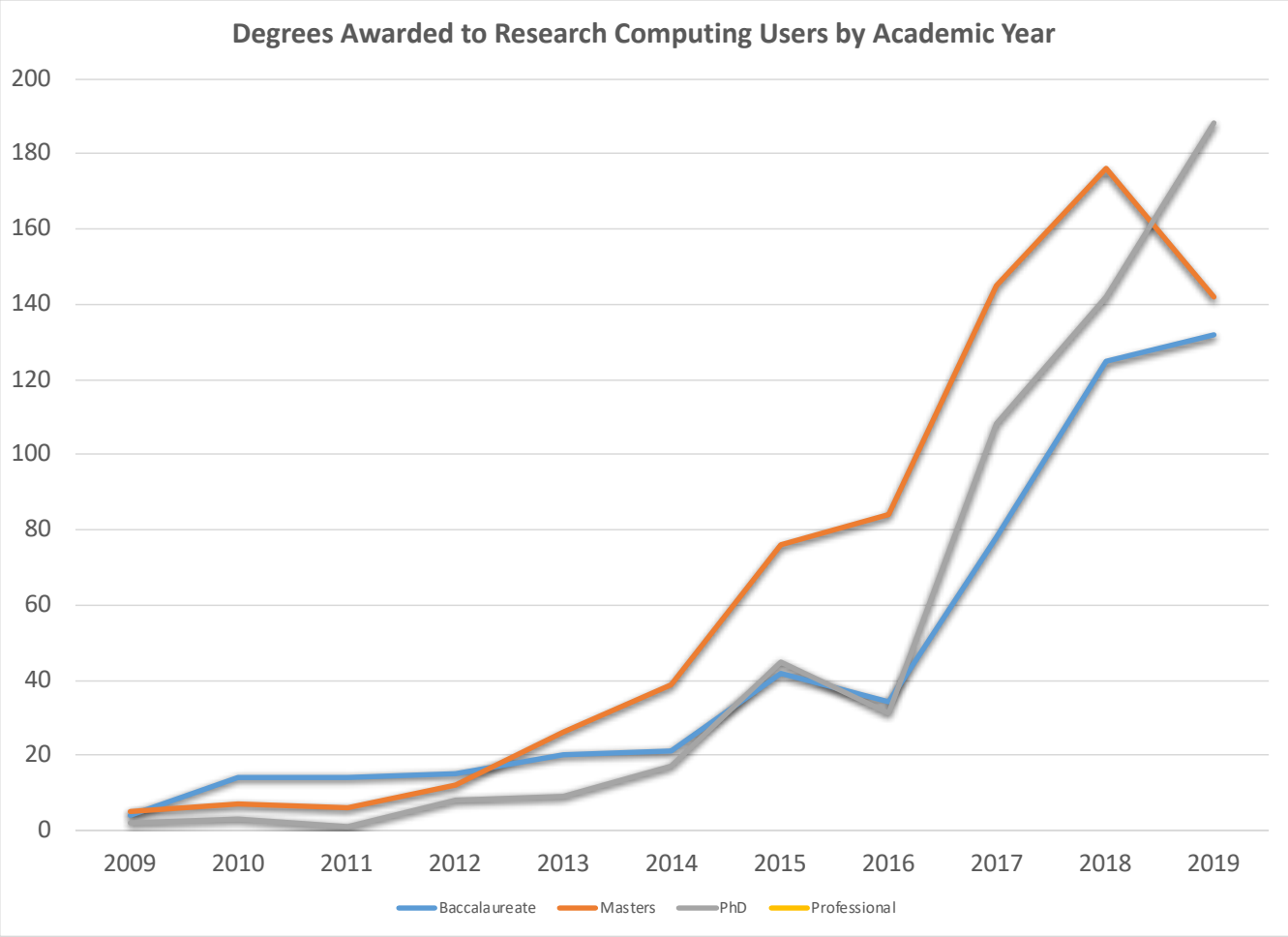
155 courses used Scholar in AY 2019-20, with over 7,000 students using the system.



PURDUE UNIVERSITY | Information Technology

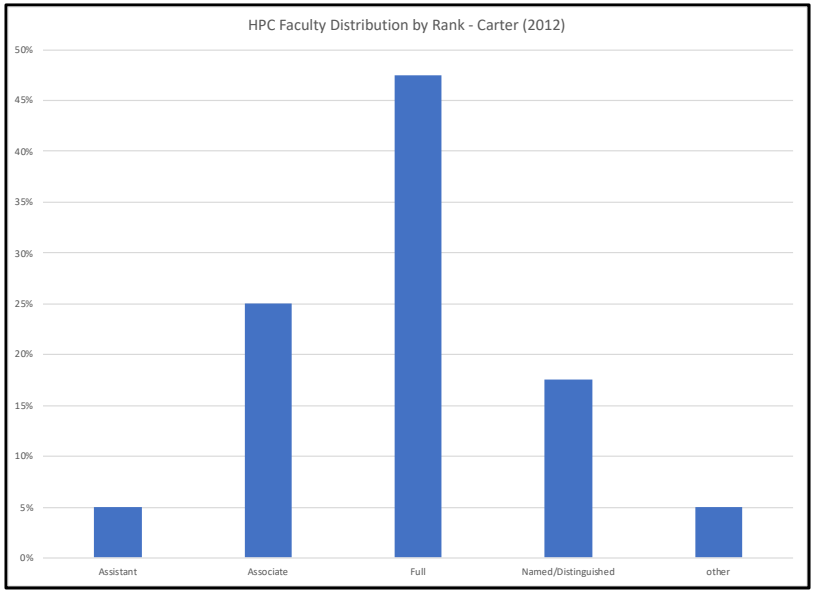
High Performance Computing is a crucial capability for STEM education and data sciences.

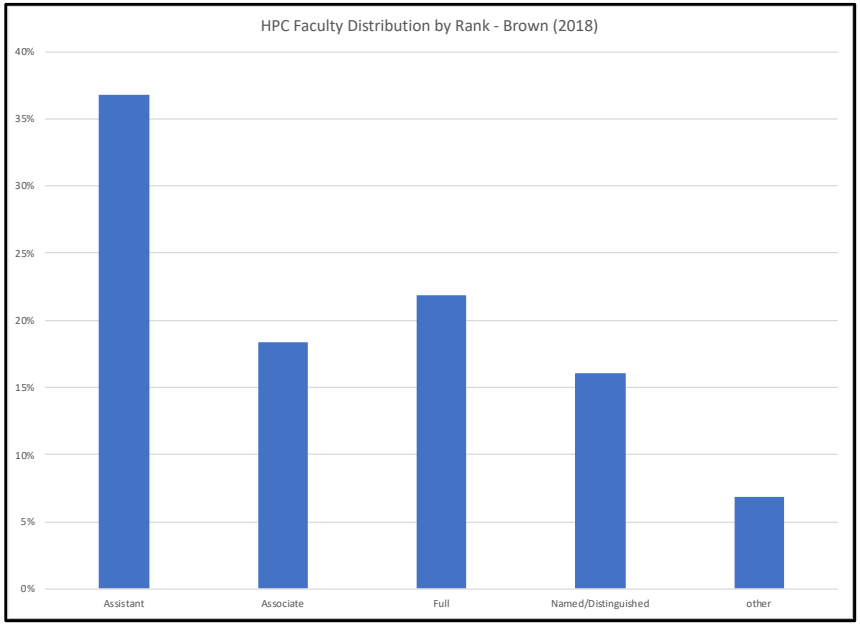
Since 2012, 2,142 degrees have been awarded to users of Purdue high-performance computing facilities.



High Performance Computing facilities are a key differentiator to recruit top faculty to Purdue

Since 2012, the share of Assistant Professor community cluster investors has grown from 5% in Carter (2012), to 38% of Brown (2017).





Since 2018, $1.1M of cluster sales came from startup accounts!

This spring alone, RCAC staff have participated in over 15 faculty interviews.

Bottom-Line - ROI

For High-Performance Computing Facilities

2020 Metric	Amount ($M)
Net Cost to Purdue	$5.88
Awards to RCAC-using Faculty	$289.90
Direct Expenditures Enabled	$242.80
F&A Expenditures Enabled	$52.70
Return on Investment	
Awards Return	49.30 x
Direct Expenditure Return	41.29 x
F&A Expenditure Return	8.96 x

Publication Output

- In CY 2019, the 267 faculty using high-performance computing since 2012 produced 1,678 peer-reviewed publications

  In that same year, research computing faculty were authors of 40% of Purdue's 57 articles in *Science, Nature,* and *PNAS*.

Journal	2019
Nature	4
Proceedings of the National Academy of Sciences of the United States of America	16
Science	3
Grand Total	**23**

Community Cluster Program

15 Years of Community Clusters

- Purdue is an early pioneer of the “condo” model.
- Cluster program started out in 2004 under first CIO Jim Bottum, with just four partners.
- Now, 215 active partners from 60 departments, from every College, and 3 Purdue campuses

353M hours delivered in 2019

CLUSTER PROGRAM PARTNERS

PURDUE UNIVERSITY | Information Technology

Department	Cores
Aeronautics and Astronautics	5740
Mechanical Engineering	5556
CMS Tier2	5440
Electrical and Computer Engineering	4344
Earth, Atmospheric, and Planetary Sciences	2540
Materials Engineering	2064
Nuclear Engineering	1564
Other College of Engineering	980
Chemistry	824
Physics and Astronomy	820
Biomedical Engineering	640
Other Executive Vice President for Research and Partnerships	600
Statistics	512
Chemical Engineering	424
Agricultural and Biological Engineering (Biological Engineering)	368
Biological Sciences	356
Industrial Engineering	296
Civil Engineering	276
Computer and Information Technology	248
Medicinal Chemistry and Molecular Pharmacology	248
Mathematics	232
Bioinformatics Core	200
Agronomy	180
ITaP	176
Computer Science	156
Horticulture and Landscape Architecture	156
Cancer Center	96
Forestry and Natural Resources	96
Biochemistry	40
Botany and Plant Pathology	40
Industrial and Physical Pharmacy	40
Brian Lamb School of Communication	32
Agricultural Economics	20
Animal Sciences	20
Food Science	20
Health Sciences	20
Other College of Pharmacy	20
Agricultural and Biological Engineering (Agricultural Systems Mgmt)	16

Community Cluster Program

Current Offerings

- Brown (Sold out, as of end of 2019!)
  - 550 nodes - Intel Xeon Gold "Sky Lake"
  - **108 unique faculty investors**
- Data Workbench
  - Interactive computing, non-batch, for data science - $300/year per lab
  - Humanities, etc.
  - 40 research labs investing to date, from 28 departments!
- GPU Clusters
  - Gilbreth – $1,600/lab annual subscription
  - 40 research labs investing to date

2018 Community Cluster – ML, AI, Data Science

Gilbreth

- GPU-based system ideal for machine learning, AI, **big data science** – as well as FEA, Chemistry, MD
- 50 nodes, 100 GPUs
  - ***>1 PF of single-precision performance!***
- 2-3PB parallel filesystem storage
- Annual subscription fee for access
- 10 new V100s added in July 2020.
- Can host custom configs!

Prof. Lillian Moller Gilbreth

Research Data Depot

Depot

Filesystem-based storage, globally accessible from HPC systems and from your laptop.

At other sites, you'll see this tier of storage described as "campaign" or "project" storage, between scratch and archive.

- 653 labs using the system

- Avg space purchase of 10 TB.
- Largest groups storing 250 TB+

Research Data Depot

Depot 2.0

- Now 6 years old, Depot 1.0 is fully allocated, and getting close to full. **~2.2 PB**
- New filesystem to carry us for the next 5 years is up and running, the bulk of labs have a first sync complete to the new filesystem. **5+ PB**
- This fall, we will schedule a cutover that we plan to be largely seamless for most labs – those of you on the top end space-wise, we will probably need to coordinate a cutover.

Controlled Computing

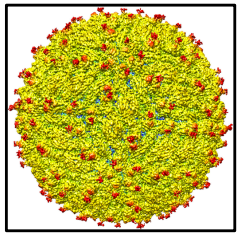
Export Controlled Computing

- All Community clusters support EAR
- "Weber" Cluster supports ITAR, CUI with NIST SP 800-171
- Purdue is well-regarded as a leader in this space

Pricing structure for Weber now more in line with community clusters – Weber is a "community cluster with enhanced cybersecurity"

Purdue Astronaut Mary Ellen Weber

Work underway for first HIPAA-aligned resource, in partnership with Regenstrief Center for Healthcare Engineering



External Funding

Anvil – NSF Supercomputing Center

- $10M capital award for acquisition of system to operate for national community via XSEDE
  - Carol Song, Preston Smith, Xiao Zhu, Rajesh Kalyanam PIs
- Plus supplemental award for staff for operations and maintenance

Potential for renewal award after 5 years

Made possible through Purdue's institutional commitment to campus computing

Anvil -System Specifications

- Partners are Dell, DDN, and Nvidia
- 1,000 nodes, based on AMD's upcoming Epyc "*Milan*" architecture
  - Direct on-chip liquid cooling!
- 5.3 PF of peak performance (will be in ~ top 50 in the world)
- 16 GPU nodes, with 4 Nvidia A100 GPUs per node
  - Total of additional 1.57 PF of GPU capacity
- 32 1 TB large memory nodes
- Composable cloud subsystem
- 100 Gbps HDR Infiniband
- 10 PB disk scratch, 3 PB flash burst buffer

CC* Award - Composable Infrastructure

Details

- Confederated Kubernetes clusters
- Software defined storage using Rook and Longhorn
- Lightweight K3s for the Edge
- Staff to assist with application deployment and scaling
- Github integration for Continuous Integration and Deployment

- Agreements with Google and Azure for cloud-native and cloud-bursting

Supported by a $392,205 award from the National Science Foundation under Grant No. **2018926***.*

Co-PIs: Hacker, Neumeister, Wisecaver, Gough

External Funding

New Projects

Awarded

- Anvil - $10M
- CC* - $400k for composable HPC (Smith, Gough, Neumeister, Hacker, Wisecaver)
- Student travel support - $10k for PEARC 20 (Hillery)
- HPC Workforce development workshop - $163K (Smith, Hacker)
- RAPID (COVID) – VR data exploration - $66k (Milisavljevic, Envision Center)

Recommended for Funding

- CCRI - $1.2M for large scale system failure research (Song, Kalyanam, Bagchi)
- Accelnet - $1.99M for global sustainability (Song, Hertel)

Purdue Projects

- RCHE – Data Analytics Platform (Hillery, Younts)
- Data Science – Krenicki Center (Lentner)
- Purdue/Lilly Endowment (Smith)
- National Defense (Ellis)

Cluster Lifecycles

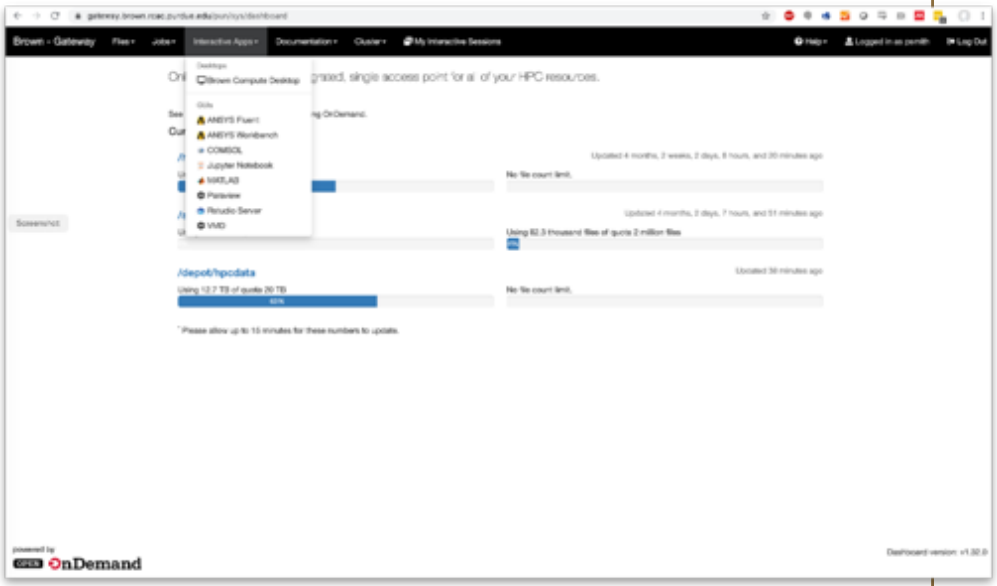
You've asked for longer...

- Rice: extended until Jan 2021 (was originally planned to retire this past May!)
- Halstead: Extended to 6 years, through Dec 2022.
- First phase of Snyder (Snyder-A/B) will retire with Rice

Value Add Features and Enhancements

Quality of Life Improvements

- Open OnDemand
  - Jupyter Notebooks
  - Rstudio
- Purdue GitHub
  - New instance up at github.itap.purdue.edu
  - Available to anybody with a Career Account, with no more request process needed!



Migration of repositories from github.rcac.purdue.edu to be coordinated

Application Software

New software stack management

- Based on “Spack”
- Will allow for automated, consistent builds across diverse platforms
- Make it scalable for us to offer a broader menu of applications

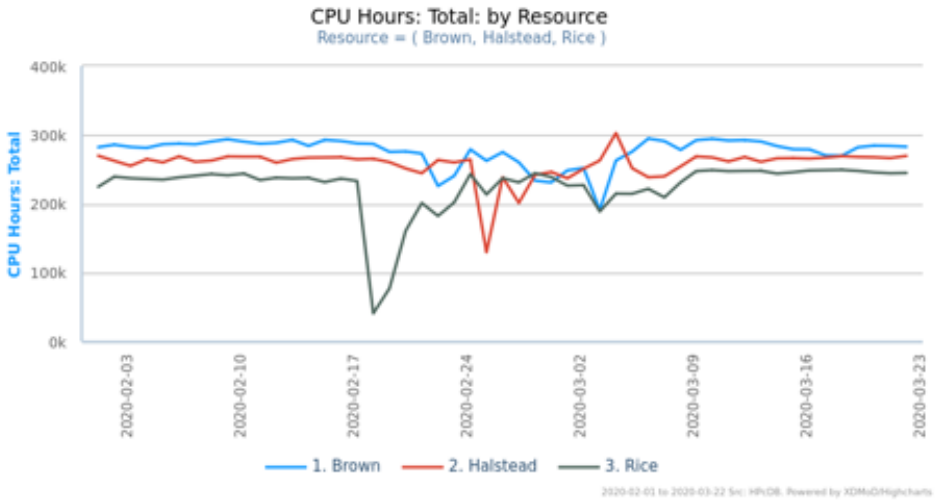
```
/bin/sh ./libtool --tag=CC   --mode=link /tmp/aai/buildgcc.tmp/./gcc/xgcc -
B/tmp/aai/buildgcc.tmp/./gcc/ -B/apps/mack/gcc/6.3.0/x86_64-pc-linux-gnu/bin/
-B/apps/mack/gcc/6.3.0/x86_64-pc-linux-gnu/lib/ -isystem
/apps/mack/gcc/6.3.0/x86_64-pc-linux-gnu/include -isystem
/apps/mack/gcc/6.3.0/x86_64-pc-linux-gnu/sys-include    -Wall -Werror  -
pthread -g -O2   -W1,-O1   -o libatomic_convenience.la  gload.lo gstore.lo
gcas.lo gexch.lo glfree.lo lock.lo init.lo fenv.lo fence.lo flag.lo
load_1_.lo store_1_.lo cas_1_.lo exch_1_.lo fadd_1_.lo fsub_1_.lo fand_1_.lo
fior_1_.lo fxor_1_.lo fnand_1_.lo tas_1_.lo load_2_.lo store_2_.lo cas_2_.lo
exch_2_.lo fadd_2_.lo fsub_2_.lo fand_2_.lo fior_2_.lo fxor_2_.lo fnand_2_.lo
tas_2_.lo load_4_.lo store_4_.lo cas_4_.lo exch_4_.lo fadd_4_.lo
fsub_4_.lo fand_4_.lo fior_4_.lo fxor_4_.lo fnand_4_.lo tas_4_.lo load_8_.lo
store_8_.lo cas_8_.lo exch_8_.lo fadd_8_.lo fsub_8_.lo fand_8_.lo fior_8_.lo
fxor_8_.lo fnand_8_.lo tas_8_.lo load_16_.lo store_16_.lo cas_16_.lo
exch_16_.lo fadd_16_.lo fsub_16_.lo fand_16_.lo fior_16_.lo fxor_16_.lo
fnand_16_.lo tas_16_.lo   load_16_1_.lo store_16_1_.lo cas_16_1_.lo
exch_16_1_.lo fadd_16_1_.lo fsub_16_1_.lo fand_16_1_.lo fior_16_1_.lo
fxor_16_1_.lo fnand_16_1_.lo tas_16_1_.lo
```

https://www.rcac.purdue.edu/news/1872

Batch System

Slurm Conversion

- After decades of being a PBS shop, we switched to Slurm this year
- Some doubts about Moab vendor's viability
- **AND** Slurm will save substantial dollars in the long run.

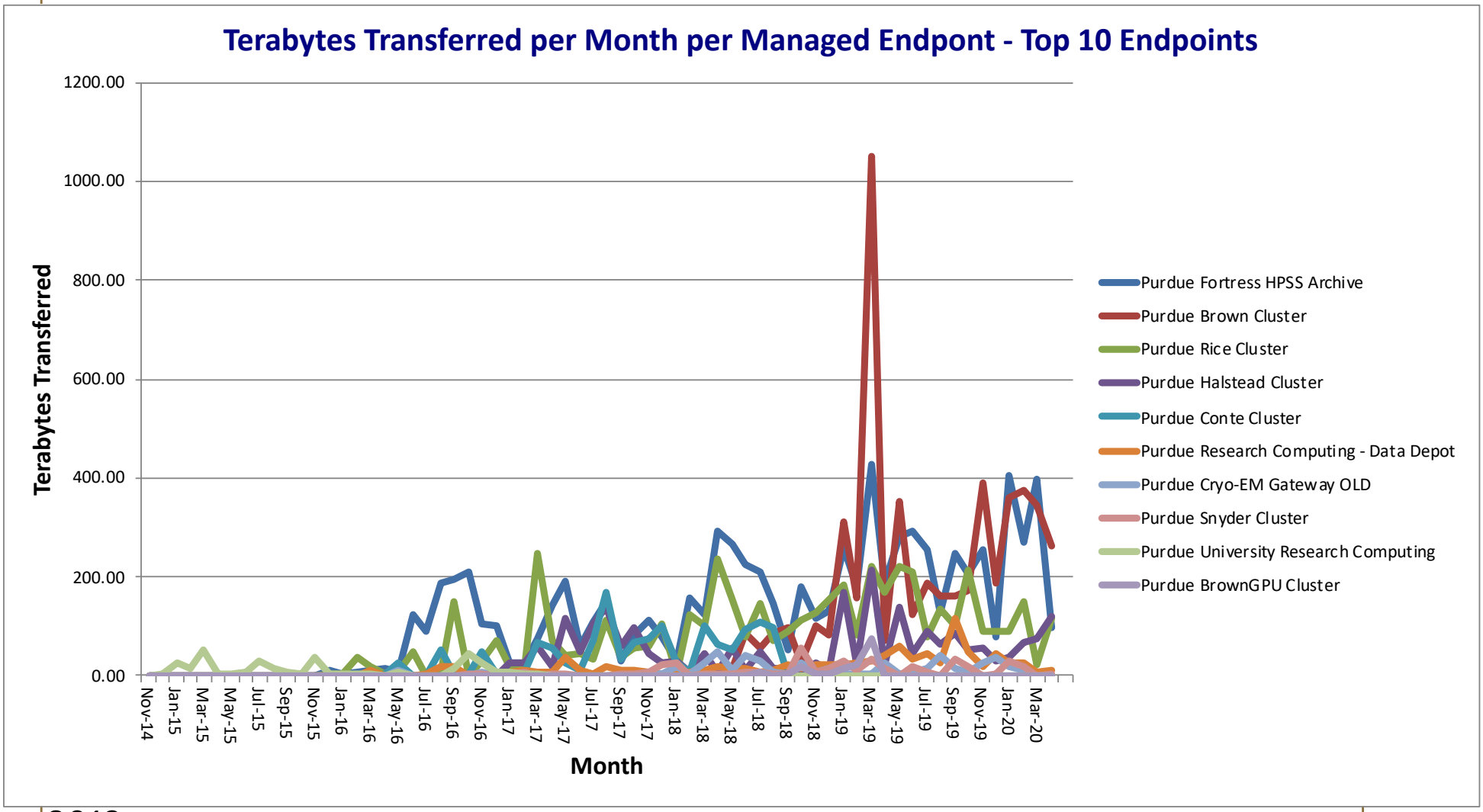


Thanks to all of you for your patience and feedback getting it deployed and tuned!

slurm workload manager

PURDUE UNIVERSITY | Information Technology

Data Transfer

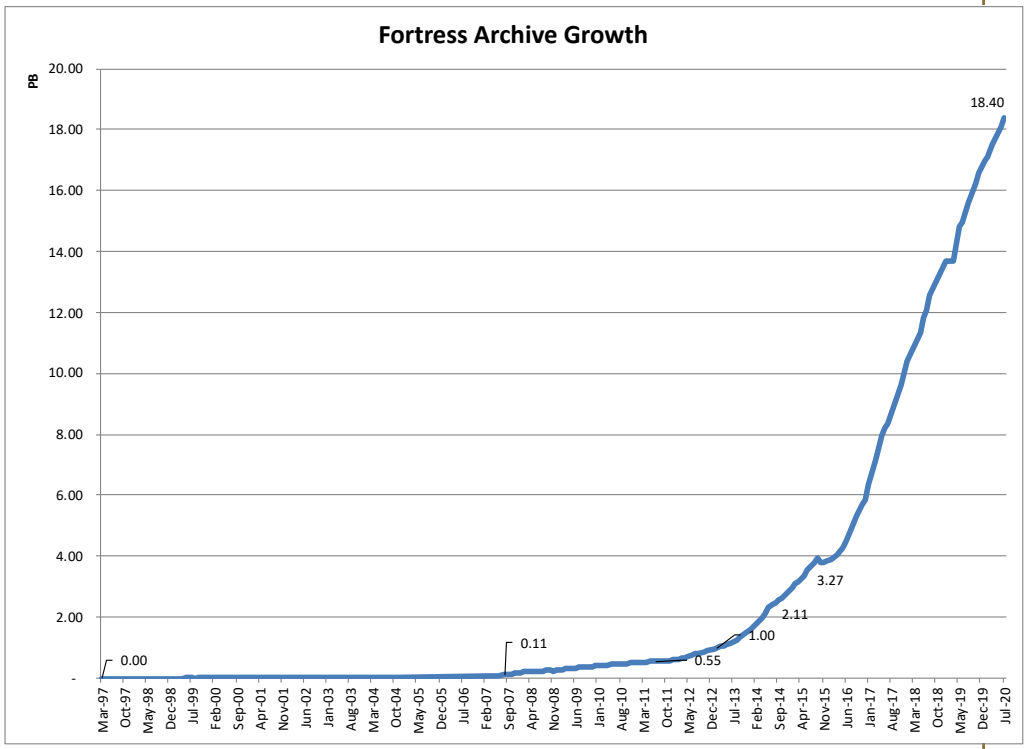


- 2019:
- 6.1 PB transferred (up from 3.1 in 2018) (> 600TB in March!)
- Average of .5 PB, 130 unique users per month

Fortress

Upgrades are Planned for FY21

- Due for tape library lifecycle upgrade and capacity expansion
- Exploring potential alternatives to HPSS
  - Cheaper
  - More functional



Cluster

2020

"Bell"

Vendor Proposals

3 Options in the Market for 2020 deployments

- Intel: Cascade Lake (basically, Brown… still)
- Intel: Cascade Lake AP
- AMD: Rome

Intel's Xeon roadmaps have failed…

- Buying a 2 year old Cascade Lake processor isn't desirable..
- Cascade Lake AP is better, but..
  - Intel's first attempt at a multi-chip processor
  - Very expensive per whole node
- AMD Rome is very compelling and is selling well in HPC market

7nm Intel CPUs are now delayed until 2022!

Cascade Lake AP

Architecture

- 2 *Skylake* dies per chip - 2x cores, but really a 4-CPU system
- 400w+ per socket!

Proposal was 112 core nodes for over $9k/node.

AMD Rome

Overview of Rome

- Faster memory (8 channels per chip @ 3200MHz)
- Simplified NUMA domain configuration (selectable)
- Centralized I/O chip to fix latency/multi-node problems seen in Naples
- Has one cycle AVX2 (50% FP improvement over Naples)

- Up to 128 cores per node!

- PCIe Gen4 for 100Gbps HDR Infiniband per node

Many HPC wins for AMD: Bell, Anvil at Purdue; Big Red 200, Jetstream 2 at IU; Frontier at ORNL; Perlmutter at NERSC; Expanse at SDSC; CERN; Campus system at Texas Tech,

11 systems on current Top 500

2020 Cluster - Bell

2020 Community Cluster - Bell

- **448 Nodes - 2.1 PF** **(2x 2013's Conte)**
- 128 core nodes(2.0 GHz base, 3.3 GHz boost), 256 GB RAM per node
- 100 Gpbs HDR Infiniband
- Direct-to-chip liquid cooling
- 5 PB Lustre Scratch (100 TB flash)
  - 50% faster than Brown's storage
- 8 Large memory nodes (1 TB RAM)
- GPU subsystem based on AMD MI50 GPUs (16 GPUs)
- Composable Kubernetes infrastructure for non-batch workflows
- **6 year cluster life**

Early User Access in Fall 2020

BELL - BENCHMARKS

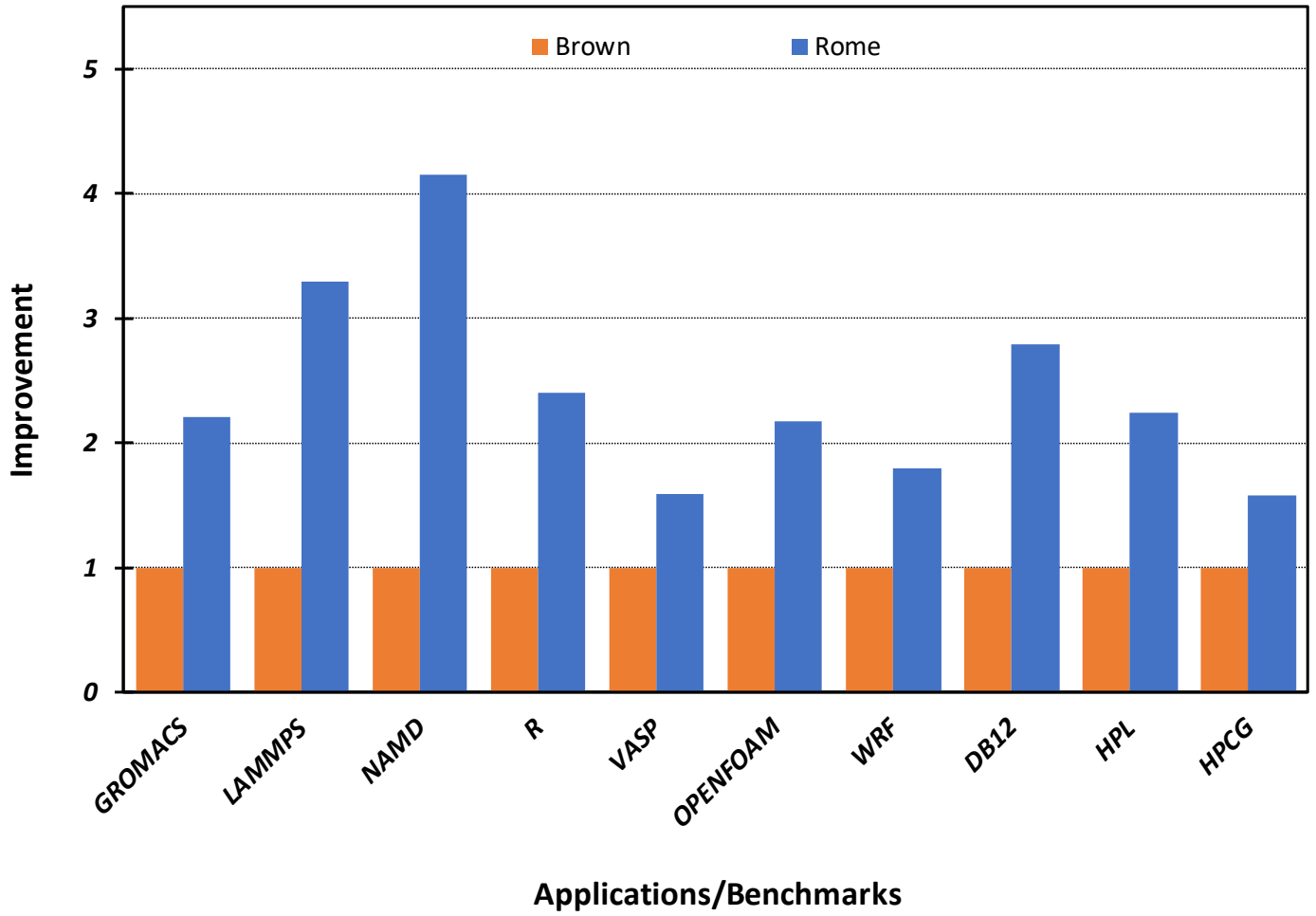
STREAM

Application		Performance on Skylake	Performance on Rome
DB12		510.0	1600.0
HPL		1518.9	3403.5
HPCG		25.5	40.2
STREAM[b]	COPY	144064.5	203522.5
	SCALE	143006.6	196341.9
	ADD	148988.5	215621.8
	TRIAD	148831.5	223098.8

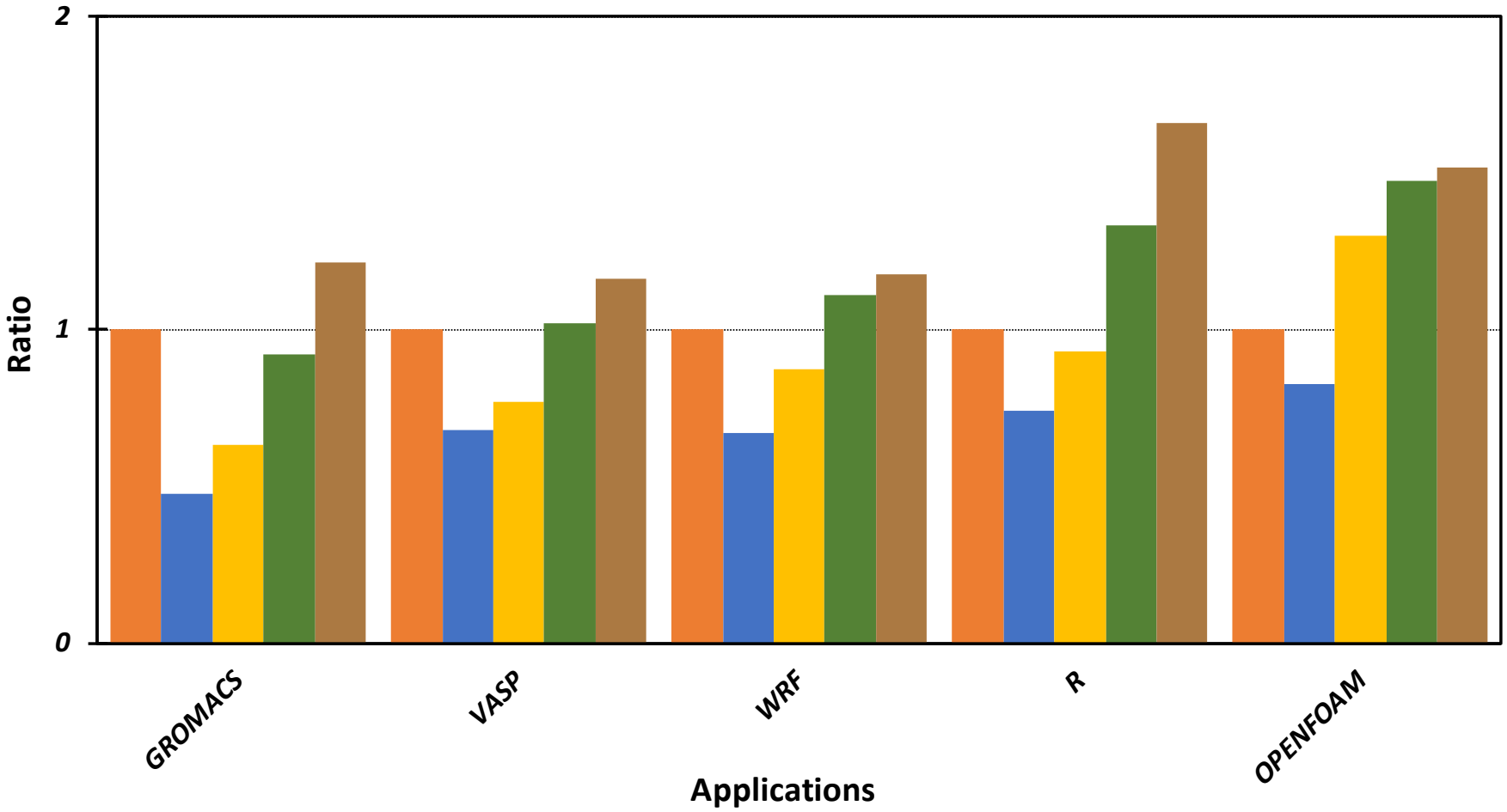
a. Unit for each DB12 entry is Monte Carlo events/s, units for each HPCG/HPL entry is GFLOPs and unit for each STREAM entry is MB/s.

b. Results when using all the cores are chosen. The all-core memory bandwidth is the more accurate representation of the memory bandwidth for majority of HPC/HTC workloads.

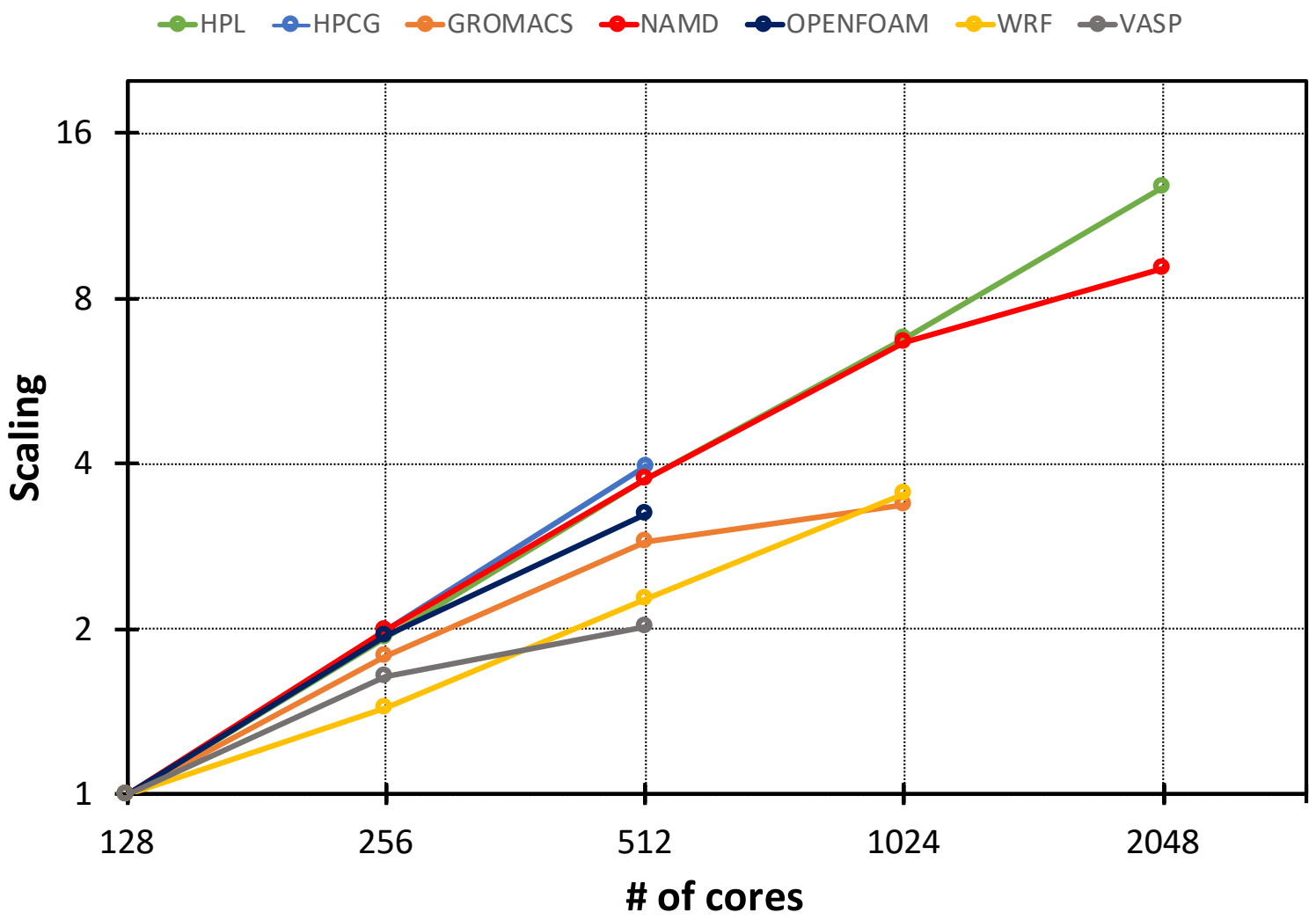
BELL -BENCHMARKS



BELL - BENCHMARKS



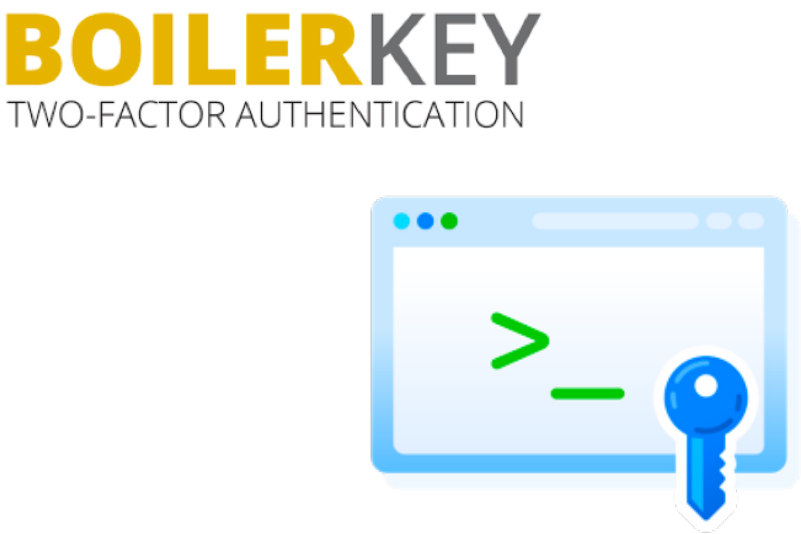
BELL - BENCHMARKS



Bell Cluster

User Experience Changes

- The 128-core Rome CPUS are the top-end, for low-end pricing! (We usually get mid-tier parts)
- Resource-specific home directory
- **Moving away from password-based auth**
  - BoilerKey two-factor
  - SSH Keys
- Specialized nodes available to all for the price of entry
  - Large memory nodes in a shared queue
  - GPU nodes in a shared queue



Bell Cluster

Costs and Details

- Price point – 64-core **shares** for $4,000
  - Compares very favorably to Halstead's 20 cores for $3,600
- Base nodes are equivalent to the bulk of Snyder
  - Snyder will not be expanded further
  - Going forward, the very-large memory need will be met by the shared queue with 1+ TB RAM nodes

Cluster	Dollars per GF	Dollars per Core
Steele	$ 34.42	$ 251.75
Coates	$ 21.83	$ 218.25
Rossmann	$ 15.25	$ 139.21
Hansen	$ 13.27	$ 122.13
Carter	$ 10.10	$ 206.25
Conte	$ 2.90	$ 418.75
Rice	$ 4.62	$ 220.00
Halstead	$ 3.78	$ 180.00
Brown	$ 2.86	$ 233.29
Bell	$ 1.57	$ 57.81

Bell Cluster

Timeline – "Fall"

- *COVID-19 has created many unknowns*
- IB Network is here
- Liquid cooling installation starts next week.
- Compute nodes in flight as well
- Scratch storage is in the datacenter now

- Once all gear arrives, and facilities are ready, integration and early access will take approx. ~4 weeks
- Then benchmark and validation

THANK YOU

rcac.purdue.edu/purchase

BENCHMARKS

Table S1(a) Single node performance from the AMD EPYC 7742 processor (Rome)[a]

Application	Performance on Skylake	Performance on Rome
GROMACS	21.5	53.4
LAMMPS	0.22	0.81
NAMD	0.96	4.48
R	12802.0	4734.0
WRF	3940.0	1952.0

•Table entries for GROMACS, LAMMPS, NAMD are in the unit of ns/day and for WRF and R are in the unit of second.

Table S1(b) Single node performance from the AMD EPYC 7742 processor (Rome)[a]

Application		Performance on Skylake	Performance on Rome
DB12		510.0	1600.0
HPL		1518.9	3403.5
HPCG		25.5	40.2
STREAM[b]	COPY	144064.5	203522.5
	SCALE	143006.6	196341.9
	ADD	148988.5	215621.8
	TRIAD	148831.5	223098.8

•Unit for each DB12 entry is Monte Carlo events/s, units for each HPCG/HPL entry is GFLOPs and unit for each STREAM entry is MB/s.
•Results when using all the cores are chosen. The all-core memory bandwidth is the more accurate representation of the memory bandwidth for majority of HPC/HTC workloads.

BENCHMARKS

Table A5 Performance data of LAMMPS on both Intel Xeon GOLD 6126 processor (Skylake) and AMD EPYC 7742 processor (Rome)[a]

# of tasks	Skylake	# of nodes On Skylake	Rome	# of nodes On Rome
1	0.73	1	0.49	1
4	2.61	1	1.85	1
8	4.67	1	3.65	1
24	8.30	1	10.06	1
48	14.75	2	18.71	1
96	24.46	4	30.06	1
128			34.64	1
256			50.53	2

•This benchmark is performed with one task per cores for the SMALL system; application performance data are in the units of ns/day.

Table A1 Performance data of DB12 on both Intel Xeon GOLD 6126 processor (Skylake) and AMD EPYC 7742 processor (Rome)[a]

Node Type	Hyper-threading	# of threads	Throughput per Node	Throughput per Thread
Rome	Off	128	1600	12.5
Skylake	Off	24	510	21.3
Skylake	On	48	686	14.3

•Application performance data are in the units of Monte Carlo events/s.