Cost and Performance Modeling for Earth System Data Management and Beyond

Jakob Luettgau, Julian Kunkel

Deutsches Klimarechenzentrum GmbH
University of Reading

HPC-IODC 2018

esiwace
CENTRE OF EXCELLENCE IN SIMULATION OF WEATHER AND CLIMATE IN EUROPE

DKRZ
DEUTSCHES KLIMARECHENZENTRUM

Outline

1. Motivation
2. Models
3. Building Blocks
4. Scenarios
5. Standards?
6. Summary

Disclaimer: This material reflects only the author's view and the EU-Commission is not responsible for any use that may be made of the information it contains

Challenges in Managing HPC Storage

- Applications create significantly more data
  - Climate/Weather: higher model resolution and larger ensembles
  - More observational data from different sources
- Increasingly heterogeneous storage landscape
  - Recent technological innovations
    - NVM, HBM, Burst Buffers, ...
  - Provides cost saving opportunities
  - Induces costs to adapt architecture, interfaces and applications
  - How to identify good configurations?
- Modeling and communication of system setup & behavior
  - Useful to conduct what-if studies
  - But no standard available
    - Every HPC site has its own representation of their HPC cluster
  - Users have difficulties to understand the model's implications
  - Graph based models were successfully applied in industry

Research Approach

Research question

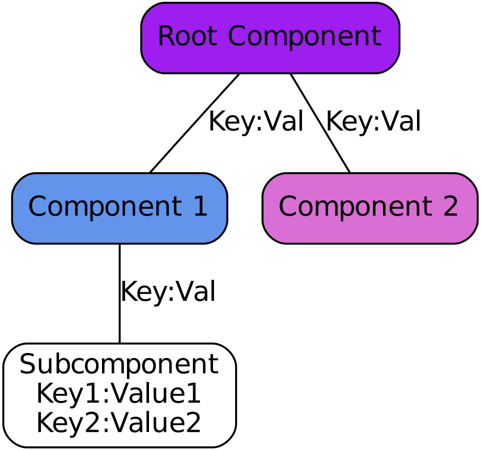
- Can we use a graphical and graph model for:
  - Modeling system topology
  - I/O path
  - Performance behavior
  - Resilience
  - Costs, Power
- Is there a way to standardize and communicate such models?

Approach

- Develop example models with trees/graphs
- Explore models for alternative DKRZ system configurations

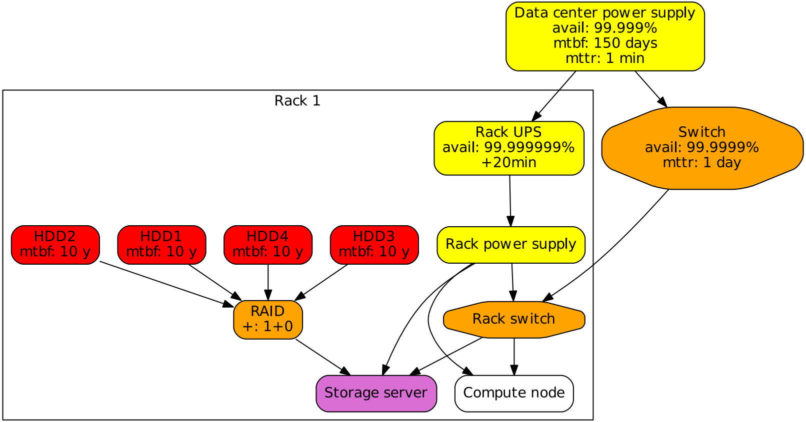
Hierarchical Model

High-level for practicality, but flexible to narrow down details where necessary.



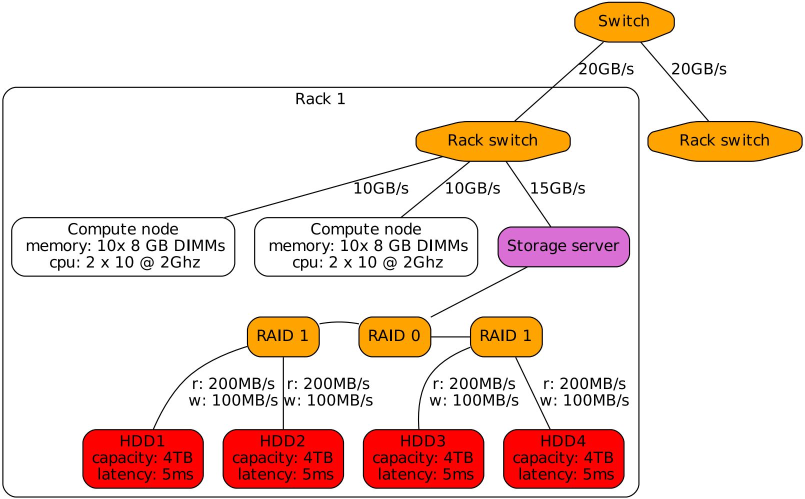
Example: Resilience

Hierarchical model including different levels of detail.



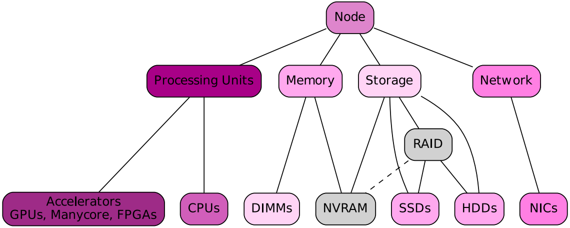
Example: Performance

Hierarchical model including different levels of detail.



Building Blocks: Nodes

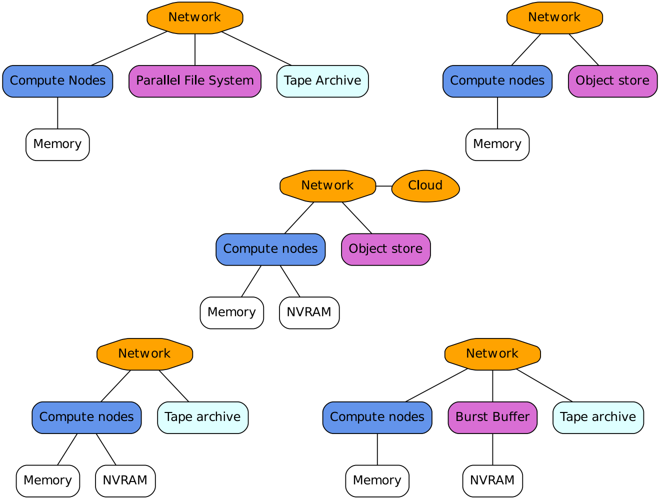
Attributing Cost or Power Consumption by Sub-Component.



- Costs/Power of a sub-component is attributed to parent
  - Actual Key/Value annotation with costs is not shown
  - Color (e.g., darkness) could encode the importance!
- Problem: sharing of functionality like NVRAM
- Some components may have alternative access path like RAID

Deployment Modes: Alternative Storage Landscapes

Just a brief glimpse of some possible data center layouts



Alternative Scenarios

Goal

- Investigate alternative storage landscapes for same investment
  - Reduce storage in favor of compute
  - Switch to object storage in favor of compute
- Assumptions: object storage costs 1/2 of PFS (vendor hint)
- Use a tabular representation
  - Often vendors have spreadsheets to make such calculations
  - Use factors to indicate how much more/less
  - Unchanged values are from the default
- Is it harder to see how costs and performance come together?

Example Scenarios for Mistral at DKRZ

	Mistral	Scale-down PFS, spent leftovers on compute		Switch to Object Storage, leftovers spent on compute	
Characteristics	Value	Factor	New value	Factor	New Value
Performance	3.1 PF/s	1.17	3.6 PF/s	1.19	3.7 PF/s
Nodes	2882	1.17	3370	1.19	3430
Node performance	1.0 TF/s				
System memory	200 TB	1.17	234 TB	1.19	238 TB
Network links	3100	1.12	3450	1.15	3565
Storage capacity	52 PB	0.5	26 PB	0.9	47 PB
Storage throughput	700 GB/s	0.5	350 GB/s	0.375	262 GB/s
Storage servers	130	0.5	65	0.75	98
Disk drives	10600	0.5	5300	0.74	7800
Archive capacity	500 PB				
Archive throughput	18 GB/s				
Compute costs	15.75 M EUR	1.17	19.53 M EUR	1.24	19.53 M EUR
Network costs	5.25 M EUR	1.10	6.04 M EUR	0.98	5.15 M EUR
Storage costs	7.5 M EUR	0.5	3.75 M EUR	0.5	3.75 M EUR
Archive costs	5 M EUR				
Building costs	5 M EUR				
Investment	38.5 M EUR		38.41 EUR		38.43 M EUR
Compute power	1100 kW	1.19	1290 kW	1.10	1309 kW
Network power	50 kW				
Storage power	250 kW	0.5	125 kW	0.75	188 kW
Archive power	25 kW				
Power consumption	1.20 MW		1.49 MW		1.57 MW

Case-study comparing Mistral as installed to a deployment with a reduced disk system and a deployment using object storage instead of a file system.

Activity: Comprehensive Data Center List (CDCL)

Contains characteristics for sites, supercomputer, and storage
`https://www.vi4io.org/hpsl/start`

System Model

- Hierarchical system model
  - Now based on an extensible JSON schema, optimized editor
  - Supports logical components and subcomponents
- Characteristics and peak values
- Measured values like Top-500

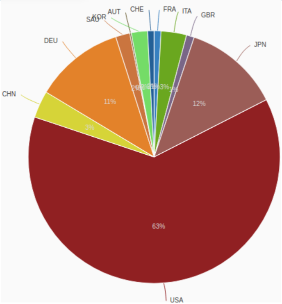
Components with characteristics

- Site, supercomputer, online storage, tape archives
- Compute nodes, storage nodes, local storage, accelerators, ...
- Supporting: e.g., CPU type, memory available, ...

CDCL Storage View 2018

Features

- Table view with selectable columns
- Flexible metrics selection/aggregation
- Multi-year analysis will be supported



Capacity grouped by country

2018

#	site.institution	site.storage system.net capacity	site.supercomputer.compute peak	site.supercomputer.memory capacity
		in PiB	in PFLOPS	in TB
1	Oak Ridge National Laboratory	250.04	220.64	3511.66
2	Los Alamos National Laboratory	72.83	11.08	2110.00
3	German Climate Computing Center	52.00	3.69	683.60
4	Lawrence Livermore National Laboratory	48.85	20.10	1500.00
5	RIKEN Advanced Institute for Computational Science	39.77	10.62	1250.00
6	National Center for Atmospheric Research	37.00	5.33	202.75
7	National Energy Research Scientific Computing Center	30.00	4.90	224.30
8	National Center for Supercomputing Applications	27.60	13.40	1649.27
9	Global Scientific Information and Computing Center	25.84	17.89	275.98
10	Joint Center for Advanced HPC	24.10	24.91	919.29
11	Cineca	23.71	12.93	455.17
12	Argonne National Laboratory	21.32	10.00	768.00
13	Forschungszentrum Jülich	20.30	6.25	454.15
14	Japan Agency for Marine-Earth Science and Technology	19.62	1.31	320.00
15	Korea Meteorological Administration	19.27	2.90	0.00
16	National Supercomputing Center in Wuxi	17.76	125.00	1310.00
17	Maryland Advanced Research Computing Center	17.00	0.87	92.67
18	King Abdullah University of Science and Technology	16.96	7.20	790.00
19	Air Force Research Laboratory	15.54	5.61	447.00
20	Leibniz Supercomputing Centre	15.00	3.58	194.00
21	National Supercomputing Center in Guangzhou	14.40	59.60	1286.00
22	National Aeronautics and Space Administration	14.21	4.97	664.00
23	Texas Advanced Computing Center	12.43	9.60	270.00
24	Engineer Research and Development Center - US Army Corps	10.66	4.57	441.60
25	Sandia National Laboratories	9.93	0.50	22.10
26	Karlsruhe Institute of Technology (KIT)	9.57	1.61	222.00
27	High-Performance Computing Centre Stuttgart	8.88	7.40	964.00
28	Total Exploration Production	8.17	6.71	54.00
29	Swiss National Supercomputing Centre	7.73	25.32	521.00
30	Eni S.p.A.	6.66	4.60	0.00
31	Nagoya University	5.33	3.20	92.00
32	PGS	5.33	5.37	584.00
33	European Centre for Medium-Range Weather Forecasts	5.33	4.25	0.00
34	Army Research Laboratory DoD Supercomputing Resource	4.09	3.70	424.00
35	University of Edinburgh	3.91	2.55	0.00
36	Pacific Northwest National Laboratory	2.40	3.40	184.00
37	Navy DoD Supercomputer Resource Center	2.11	2.05	0.00
38	Vienna Scientific Cluster	1.81	0.68	42.18
39	Center for Scientific Computing	0.75	0.51	77.57

A Little Bit Towards Standards

- Towards Javascript for embedding into a data center web page
- Allowing the site to describe and visualize their system
- Hosted by the site directly
- Allowing a simple export into VI4IO data center list

$\Rightarrow$ Towards a **standardized presentation** of systems !

Summary, Status and Outlook

- Graph modeling is useful for various reasons
  - Approach balances practicality with opportunity to add details
- There exists no appropriate standard
- Case-study for alternative storage landscapes for Mistral
- We believe standards accelerate comparison and analysis

Outlook

- Standardized description for HPC users, system developers
- Machine readable specifications
- Compatible with modeling tools of community
- Specifications of components should be provided by vendors

We welcome interest in standardization around this topic!

The ESiWACE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No **675191**

CENTRE OF EXCELLENCE IN SIMULATION OF WEATHER AND CLIMATE IN EUROPE

Disclaimer: This material reflects only the author's view and the EU-Commission is not responsible for any use that may be made of the information it contains

Integration with Earth System Data Middleware

Adaptively choose backends. Discriminate by data, metadata and access type.