Exceptional service in the national interest

SAND2019-12953PE

Improving HPC Productivity Through Monitoring, Analysis, and Feedback

Jim Brandt (09328)

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Meet The Team

- **SNL:** Omar Aaziz, Ben Allan, Jim Brandt, Ann Gentile, Ben Schwaller
- **Open Grid Computing:** Nichamon Naksinehaboon, Narate Taerat, Nick Tucker, Tom Tucker
- **BU:** Burak Aksar, Emre Ates, Prof. Ayse Coskun
- **NMSU:** Prof. Jon Cook
- **UIUC:** Saurabh Jha, Archit Patke
- **UCF:** Prof. Damian Dechev, Ramin Izadpanah
- **NU:** Prof. Devesh Tiwari, Tirthak Patel

Sandia Production HPC Consistently Runs at 100% of Capacity

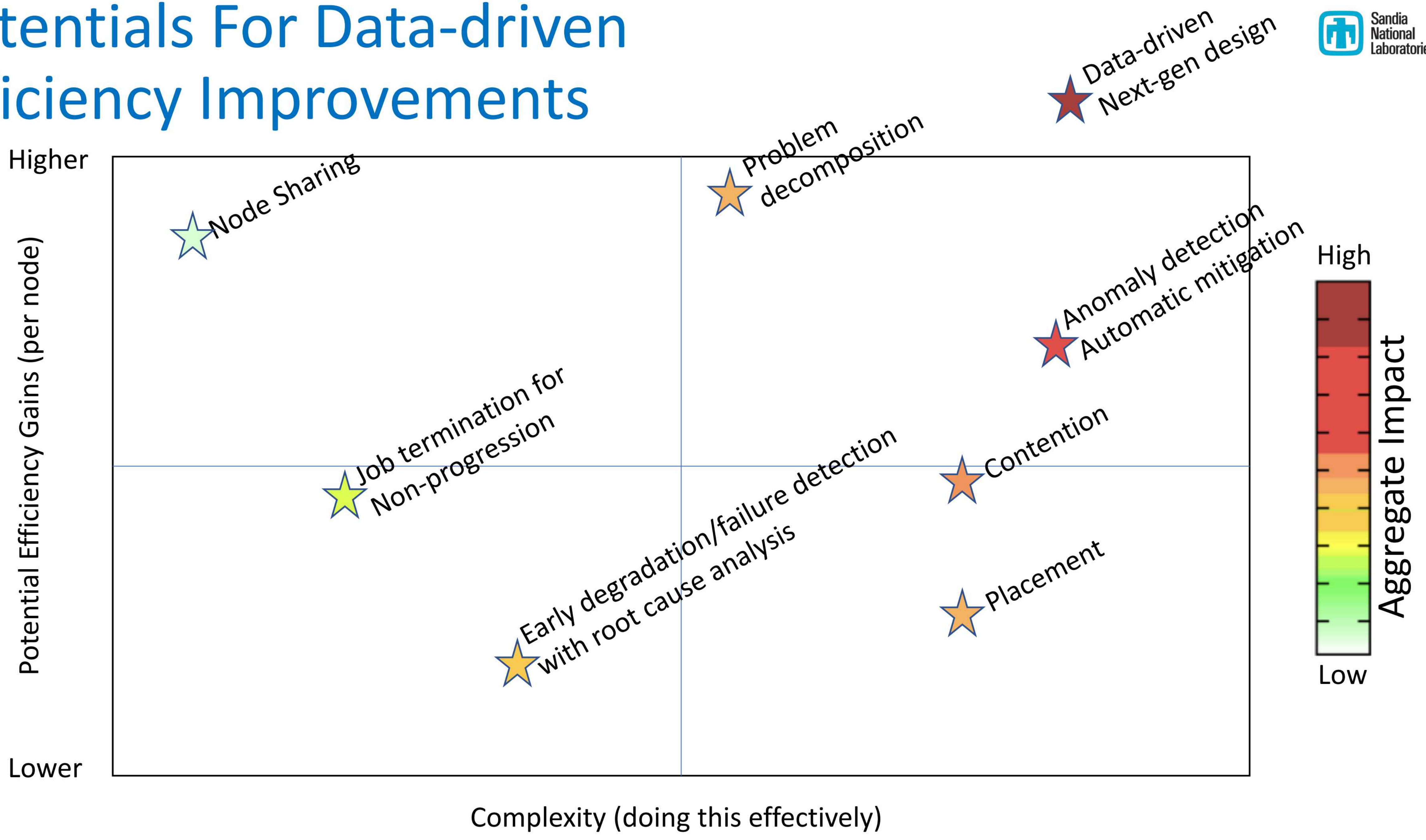
Sandia National Laboratories

System	Vendor	Nodes	Processor cores total	Processor:Sockets:Cores/Socket	Memory/ Node	Memory / Core	TFLOPS (peak)	Processor Hours/Year
				Available 10/2019				
Attaway	Penguin	1,488	53,568	2.3 GHz Intel Skylake 2SL:18C	192	5.33	1,800	469,255,680
Serrano	Penguin	1,122	40,392	2.1 GHz Intel Broadwell:2S:18C	128	3.56	1,357	353,833,920
Sky Bridge	Cray	1,848	29,568	2.6 GHz Intel Sandy Bridge:2S:8C	64	4.00	588	259,015,680
Chama	Appro	1,232	19,712	2.6 GHz Intel Sandy Bridge:2S:8C	64	4.00	392	172,677,120
Uno	Dell	251	3,344	2.7 GHz Intel Sandy Bridge:2S:8C/4S:8C	64/128	4/8	71	29,293,440
Eclipse	Penguin	1,488	53,568	2.1 GHz Intel Broadwell:2S:18C	128	3.56	1,800	469,255,680
Doom (GPU)	*Penguin*	*30*	*431,160*	*2.1 GHz Intel Broadwell:2S:18C 4X Nvidia P100 GPUs/node*	*512*	*14.22*	*565*	*3,776,961,600*
Ghost	Penguin	740	26,640	2.1 GHz Intel Broadwell:2S:18C	128	3.56	895	233,366,400
Black Total:		6,711	604,384				5,668	5,294,403,840
Cayenne	Penguin	1,122	40,392	2.1 GHz Intel Broadwell:2S:18C	128	3.56	1,357	353,833,920
Jemez	HP	288	4,608	2.6 GHz Intel Sandy Bridge:2S:8C	32	2.00	95	40,366,080
Pecos	Appro	1,232	19,712	2.6 GHz Intel Sandy Bridge:2S:8C	64	4.00	392	172,677,120
Red Total:		2,642	64,712				1,844	566,877,120
Solo	*Penguin*	*374*	*13464*	*2.1 GHz Intel Broadwell:2S:18C*	*128*	*3.56*	*452*	*117,944,640*
Green Total:		374	13,464				452	117,944,640
Bridge	Appro	1,848	29,568	2.6 GHz Intel Sandy Bridge:2S:8C	64	4.00	588	259,015,680
Sand	Appro	924	14,784	2.6 GHz Intel Sandy Bridge:2S:8C	64	4	294	129,507,840
Orange Total:		2,772	44,352				882	388,523,520
TOTALS:								
Traditional CPU's:		12,469	295,752				8,282	2,590,787,520

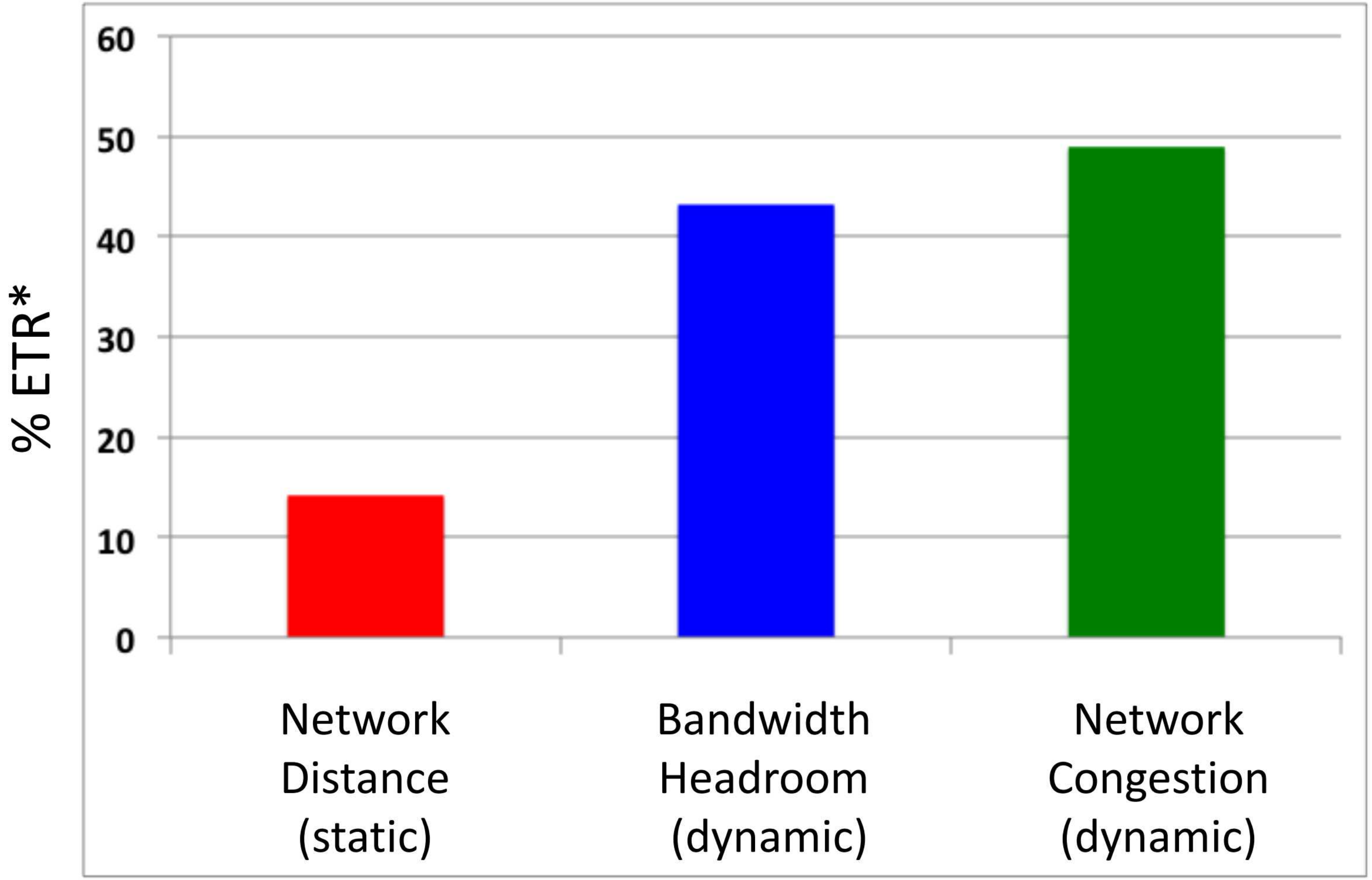
Black Multi-Cluster file systems:
16 PB Lustre (/gscratch,/nscratch)
4 PB GPFS (/gpfs1)
70 TB NAS (home and projects)

Red Multi-Cluster file systems:
16 PB Lustre (iscratch,mscratch)
4 PB GPFS (/gpfs1)
70 TB NAS (home and projects)

Potentials For Data-driven Efficiency Improvements



Using LDMS to Enable **Run-Time Assessment and Response**



Remapping based on **dynamic network assessment** recovered ~50% of the time otherwise lost to heavy congestion.

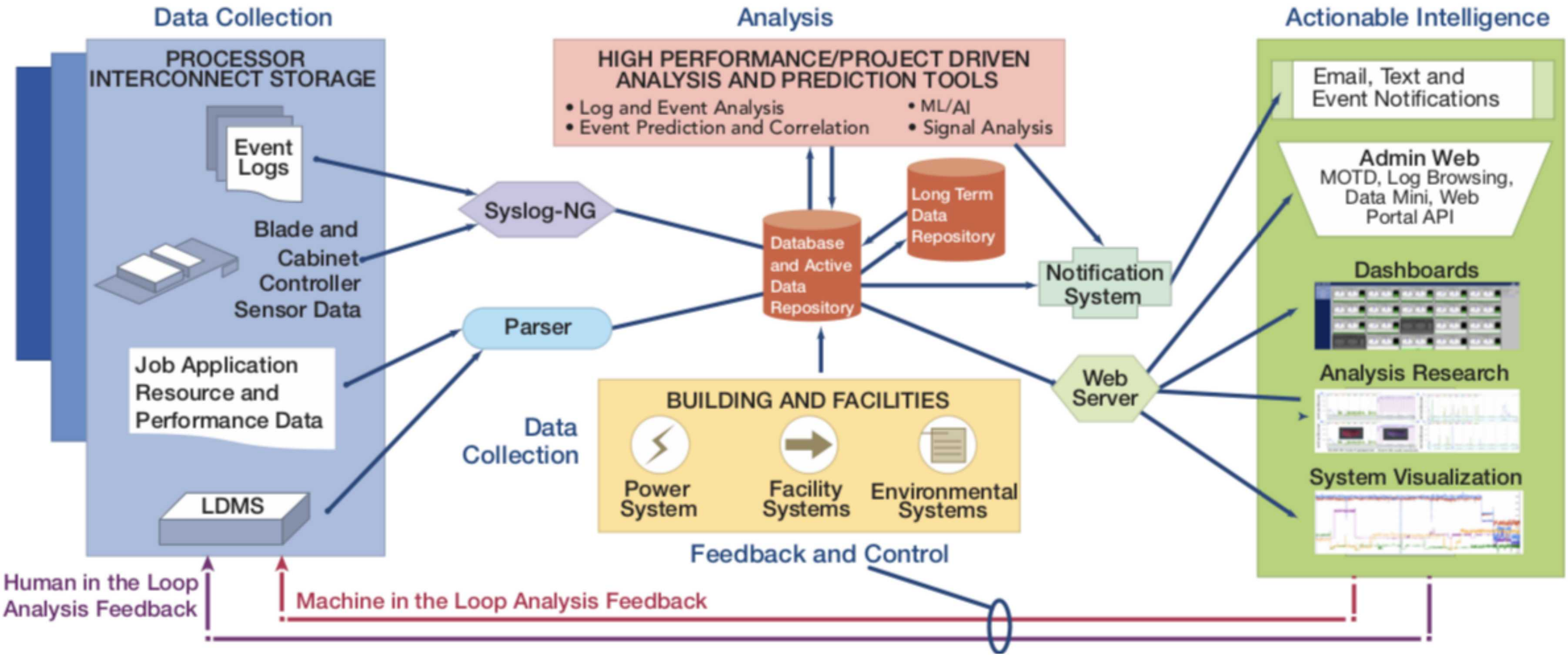
Graph edges weighted by network:

- network distance (hops)
- bandwidth headroom
- congestion (credit stalls)

*Execution **Time Recovered**
(higher is better)

From: Demonstrating Improved Application Performance Using Dynamic Monitoring and Task Mapping HPCMASPA @ IEEECluster 2014

Holistic HPC Monitoring and Analysis System Architectural Overview

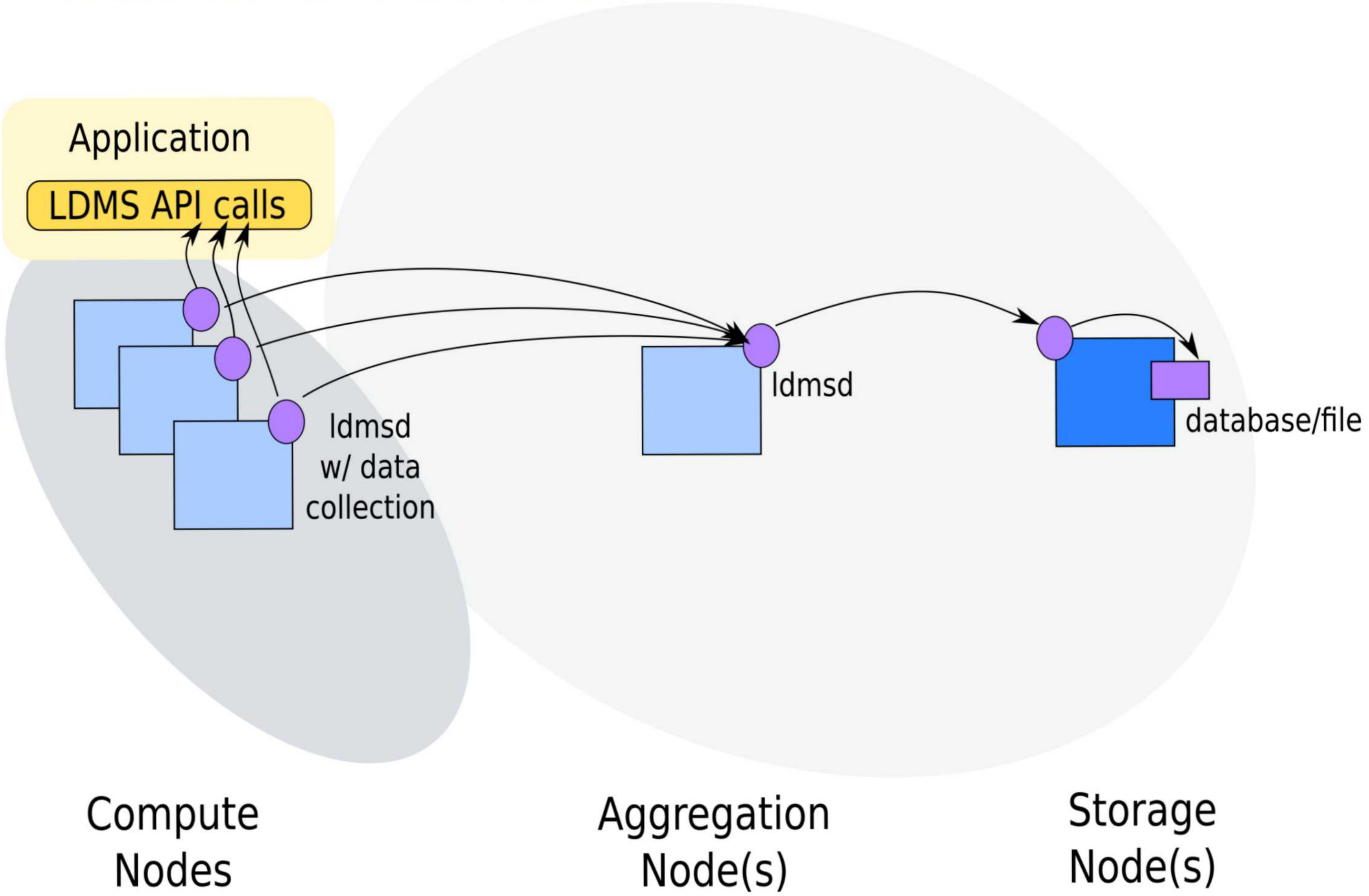


From: Holistic Measurement Driven System Assessment – *ECP Annual Meeting 2019*

Technical Challenges

- **Whole-system data collection and transport cannot impact performance**
- **Data management must efficiently support run-time and historical analysis**
- **Extraction of actionable information (analysis) from high volume (~10s of TB/day) and high dimension (100s to 1000s of discreet variables) data**
- **Relevant** raw **data** may be **un-exposed or unknown** and change from architecture-to-architecture
- How raw and derived **data** can be **quantifiably associated with application impact** may be unknown
- Explainable and interpretable Machine Learning (ML) application to **active feedback** systems
- Analytical approaches need to support incorporation of **architecture-relevant information**, and provide **confidence bounds** for response
- **Unknown** which **application** runs should have similar performance, to detect abnormal performance
- Getting **early access** to machines: "rare" events occur more frequently
- Cleansing and **releasing data** for analysis research is problematic

Lightweight Distributed Metric Service (LDMS) High Level Overview



* Only the current data is retained on-node

Differentiating LDMS Features

Feature	Benefit
Optimized Data Structures	Low Overhead (cpu, network) and Bounded Footprint
Synchronized Data Sampling	System State “Snapshot” for Coherent Analysis
Resilience	Automated Failover & Recovery of Aggregators
Large Array of Transports	Run on IB, Ethernet, OmniPath, Aries, Gemini
Plugin Based	Easy to contribute to: samplers, stores, authentication
RDMA based PUB/SUB	Low Overhead Compared With Kafka, Rabbit
On-the-Fly Sampling Interval Change	High Fidelity On Demand
Sampling and Transport Rate Independent	Low Overhead and Network Resiliency
RDMA & Socket Communication	Lightweight Over Any RDMA Enabled Transport
Job & Application Info. With Data	Simplifies Job Based Analysis
Platform Independent	Run On X86, ARM, Power
Large Fan-in	Minimal Additional Infrastructure
Security	Range of authentication plugins (e.g., munge)

Metric Set “meminfo”

```
voltrino:/opt/ovis/etc/ldms # ldms_ls -h nid00006 -x sock -p 412 -a munge -lv nid00063/meminfo
Schema         Instance                 Flags  Msize  Dsize  UID    GID    Perm       Update            Duration          Info
-------------- ------------------------ ------ ------ ------ ------ ------ ---------- ----------------- ----------------- --------
meminfo        nid00063/meminfo            CR    1976    416      0  44476 -rwxrwx--- 1570023154.002980          0.000031 "updt_hint_us"="1000000:300000"
-------------- ------------------------ ------ ------ ------ ------ ------ ---------- ----------------- ----------------- --------
Total Sets: 1, Meta Data (kB): 1.98, Data (kB) 0.42, Memory (kB): 2.39

=======================================================================

nid00063/meminfo: consistent, last update: Wed Oct 02 07:32:34 2019 -0600 [2980us]
M u64        component_id                              63
D u64        job_id                                    0
D u64        app_id                                    0
D u64        MemTotal                                  131899768
D u64        MemFree                                   129865232
D u64        MemAvailable                              129416140
D u64        Buffers                                   23380
D u64        Cached                                    508992
D u64        SwapCached                                0
D u64        Active                                    277760
D u64        Inactive                                  307892
D u64        Active(anon)                              193908
D u64        Inactive(anon)                            184852
D u64        Active(file)                              83852
```

Metric Set “syspapi”

```
voltrino:/opt/ovis/etc/ldms # ldms_ls -h nid00006 -x sock -p 412 -a munge -lv nid00063/syspapi
Schema         Instance                 Flags  Msize  Dsize  UID    GID    Perm       Update            Duration          Info
-------------- ------------------------ ------ ------ ------ ------ ------ ---------- ----------------- ----------------- --------
syspapi        nid00063/syspapi            CR     680   5704      0      0 -rwxrwxrwx 1570023691.008566          0.004697 "updt_hint_us"="1000000:300000"

-------------- ------------------------ ------ ------ ------ ------ ------ ---------- ----------------- ----------------- --------
Total Sets: 1, Meta Data (kB): 0.68, Data (kB) 5.70, Memory (kB): 6.38

=======================================================================

nid00063/syspapi: consistent, last update: Wed Oct 02 07:41:31 2019 -0600 [8566us]
M u64        component_id                               63
D u64        job_id                                     0
D u64        app_id                                     0
D u64[]      PAPI_TOT_INS                               35962599040,33992506816,35016600026,34797373446,34579431570,34600633036,34866212892,35120182994,3
4647555149,34187973522,36058949119,42688377667,44405137911,43070075781,39425173545,38523220124,82390828804,77743186601,56745278099,54416189686,4827372438
4,49570082931,48293353594,46769113539,46073901733,43311232326,43761299041,42782782577,42745101204,42900537963,43035807262,43403273118,42673441079,3915072
0271,96524480804,69094274050,102400075682,241895203414,42615199699,41434918343,42160727972,41640611193,42809823414,33640896301,32903940720,33804660746,46
749061082,45350156751,35889486313,35658688722,35233565448,43172353875,34502476895,35716981879,35572136501,34847601434,34842029631,33103620319,32842315011
,32324849160,32301258050,32580067510,33029662641,78591868588
D u64[]      PAPI_TOT_CYC                               1091371150343,1074482294242,1075856872394,1007361704285,1047474321368,1070600164757,1060578247208
,1068145172379,1065221777202,1033641916867,1046781299237,1048708645507,1056821194745,1063166574299,1273045767276,1265521218824,1145321347062,116919526242
5,1106759879593,1062340319720,1043320992729,1073142063380,1090214088094,1095190871410,1115609545140,1098615119177,1068226323764,1077484495170,10795218837
42,1077241702960,1087609948114,1033060164149,871582062019,888198426532,1102824860297,913000708653,1093591218121,1835082078673,870873545096,868985415262,8
75819215958,944906582576,958630620152,849662992559,840151846041,857964093239,763520739966,747993675182,1195050125939,1236902581229,1221504257051,11920721
00388,1140989105013,1189871177680,1193022222152,1189308106075,1219874884039,1208104501799,1144802530038,1164099423137,1161444479999,1177150316373,1182536
171084,1505220945939
D u64[]      PAPI_LD_INS                                14251682388,12749639917,13666872842,13763405180,13950309203,12647251640,13104328219,13334681962,1
3427661219,11988400650,12328489316,14364972540,14562701610,14415386254,20306398779,19679719376,32843837244,28969476391,21876024268,20358724627,1892530368
9,19105143594,18927425167,18500434180,18423431144,17942491383,18013074508,17737635374,17699555908,17814161762,17765160158,18266044502,24806214894,2529241
7452,42097753013,33221459130,43434361660,94531460611,25517034945,25029921947,25377767841,23101290622,23672534844,24200240512,23915997328,24270879669,2768
2639706,27959812405,24952895443,24939427388,24768692145,27201876895,24531202462,24918645614,12288287508,12201030582,12220899474,11917990036,11926292606,1
1829570744,11847731021,11832208938,12123376320,20919526663
~
```

Installations & **Active Contributors**

- NNSA
  - **LANL** – ATS (Trinity), CTS, and other
  - **LLNL** – CTS
  - **SNL** – ART, CTS
  - TOSS stack
- Office of Science:
  - **NERSC** – Cori, Edison
- NSF
  - NCSA – Blue Waters
- Industry
  - **Cray***
  - Exxon Mobile
- Non-US
  - AWE
  - HLRS
- Misc Universities

*Cray has Integrated LDMS into its Shasta platform monitoring solution

DOE HPC Facilities

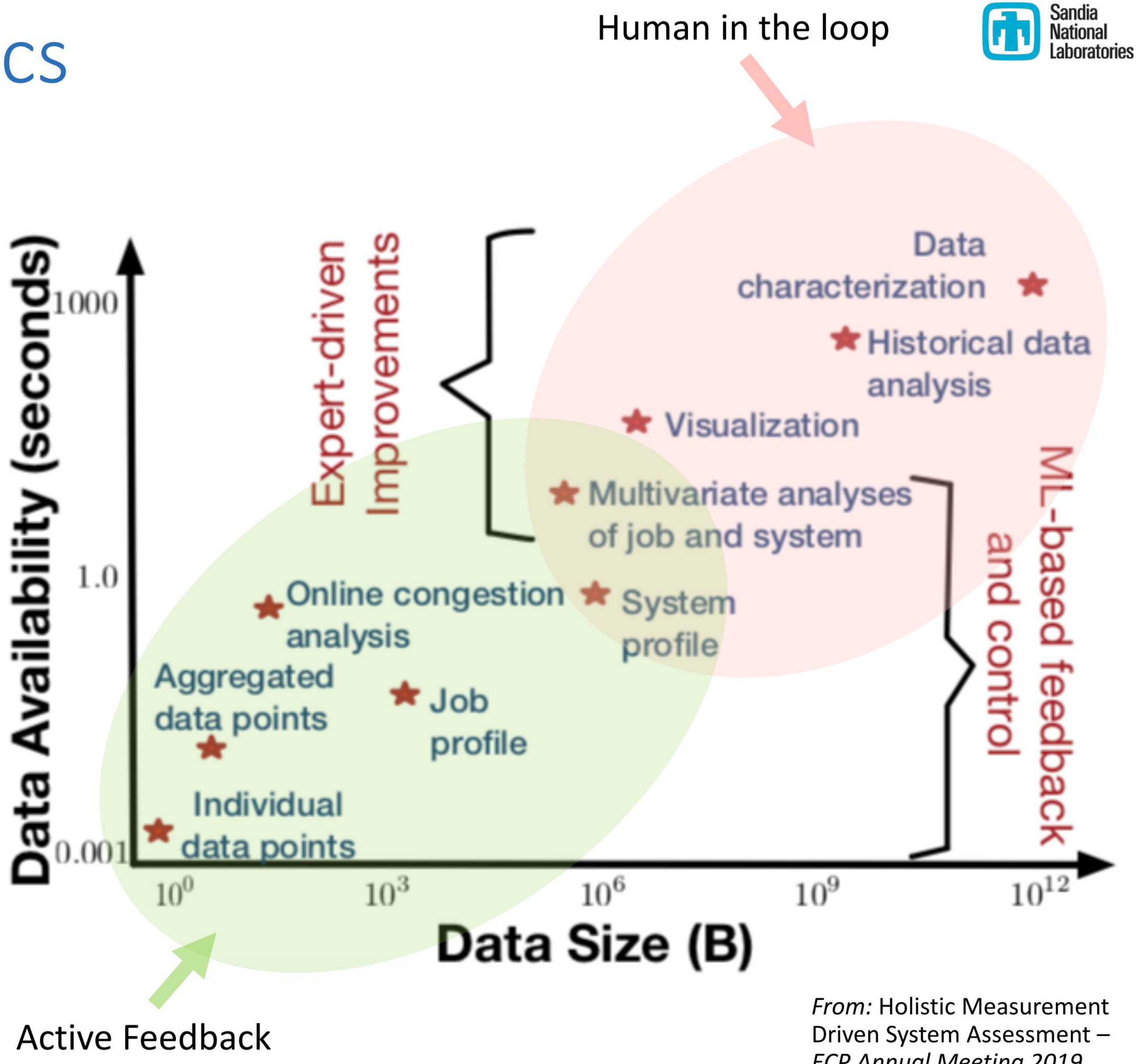
✓ -- Deployed or planned deployment

✓ -- Being evaluated for deployment

Base Source: hpcwire

Actionable Data Analytics

- Data features:
  - High volume (~10s of TB/day)
  - High dimension (100s to 1000s of discreet variables) data
  - **Getting** data is *not* a challenge!
- ASC Machine Learning (ML) focus areas:
  - Validated and explainable ML
  - Physics/boundary-constrained ML (e.g., domain-relevant)



From: Holistic Measurement Driven System Assessment – *ECP Annual Meeting 2019*

University Collaborations

Collaborations with five universities on advanced **data analytics** and **sampler development**

BOSTON UNIVERSITY — Machine Learning (ML) based anomaly detection and characterization

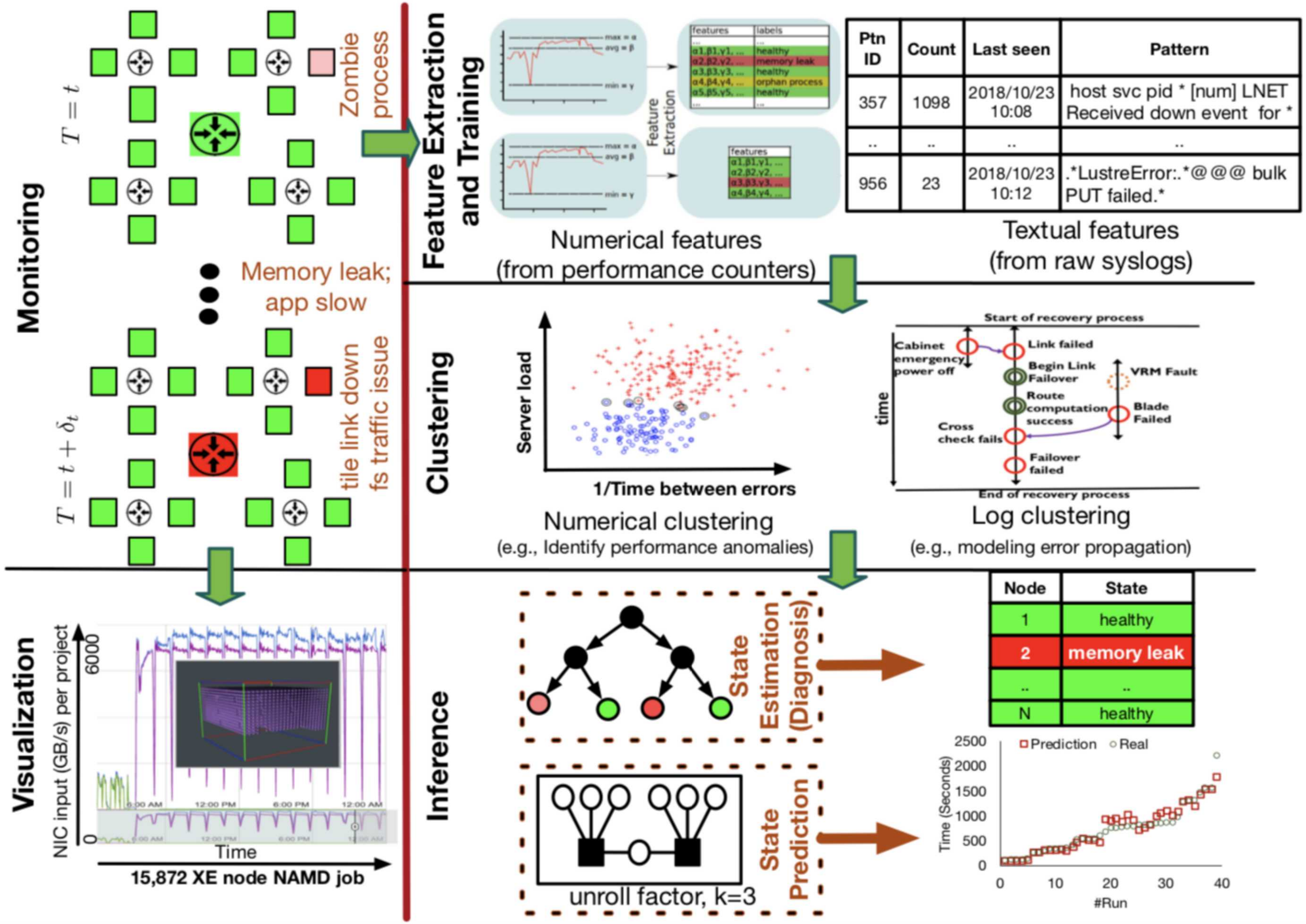
ILLINOIS — Statistics and ML based network congestion characterization and mitigation

NM STATE UNIVERSITY — Application characterization using application instrumentation and run-time analysis

UNIVERSITY OF CENTRAL FLORIDA 1963 — MPI I/O monitoring and application network resource characterization

LVX VERITAS VIRTVS — Characterization and mitigation of network congestion effects using application instrumentation

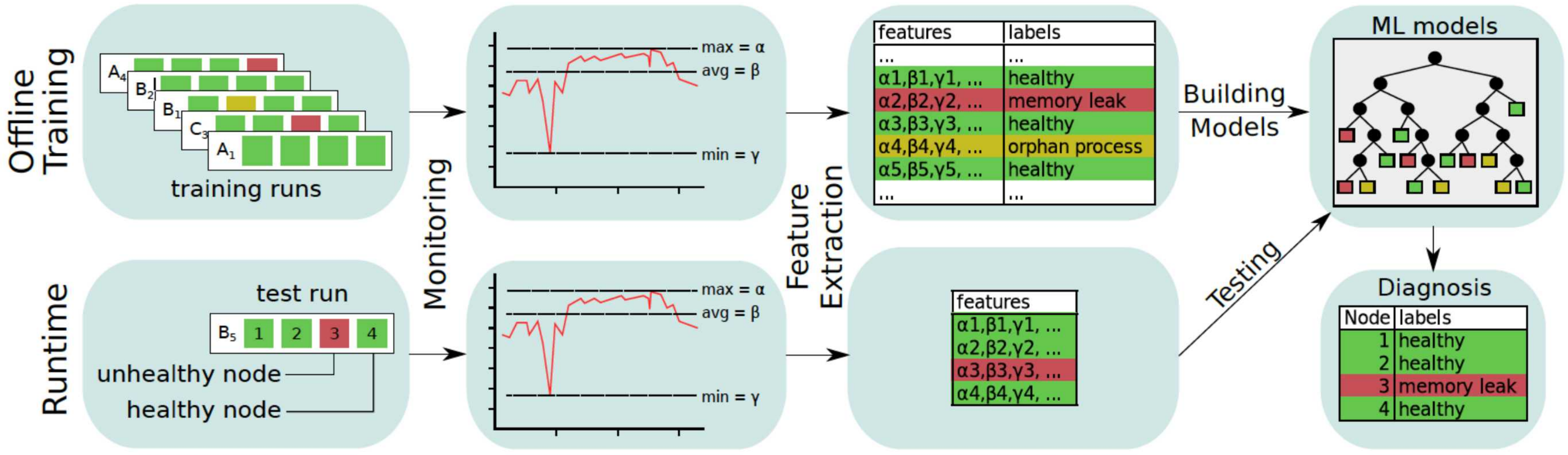
Analytics and Feedback (High Level)



From: Holistic Measurement Driven System Assessment – *ECP Annual Meeting 2019*

Anomaly Detection and Problem Diagnosis

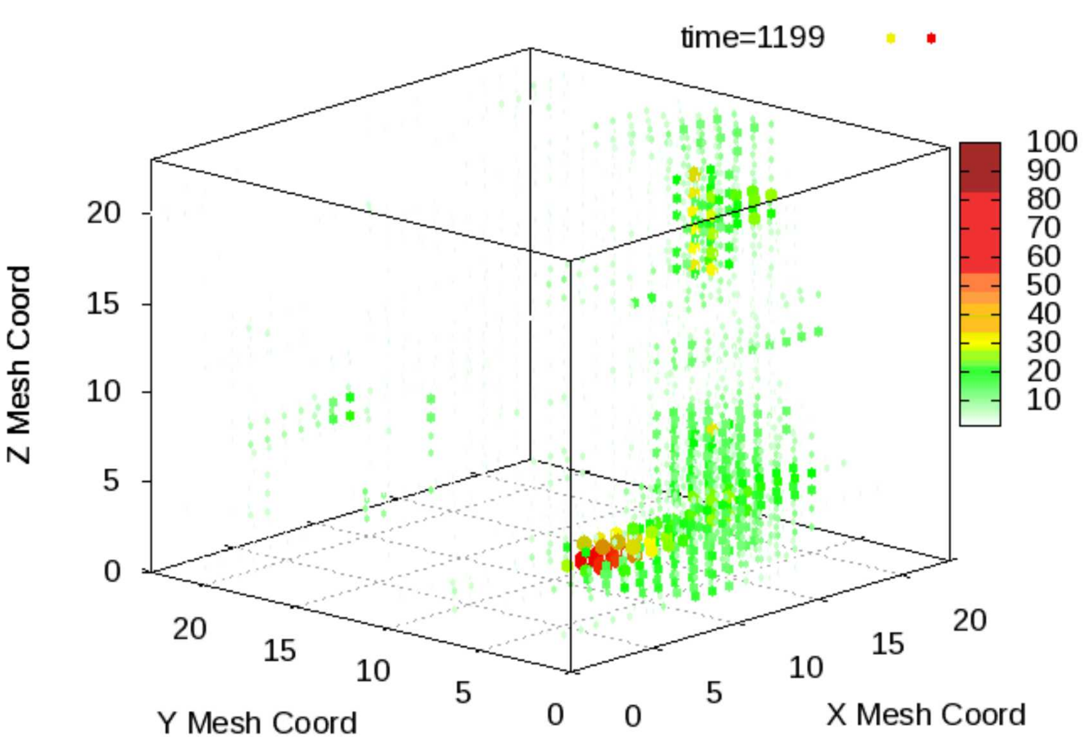
- Analysis: Use Machine Learning (ML) to build models of **normal and abnormal behavior** for each condition. Detect and diagnose run time performance/security problems through classifying data with respect to models.
- **Detect and diagnose** performance issues associated with learned **behavioral characteristics**: memory leaks, network contention, I/O contention, rogue processes, etc.



From: Diagnosing Performance Variations in HPC Applications Using Machine Learning ISC 2017 -- Gauss Award Winner

Machine learning models built offline are used for classifying observations and runtime assessment

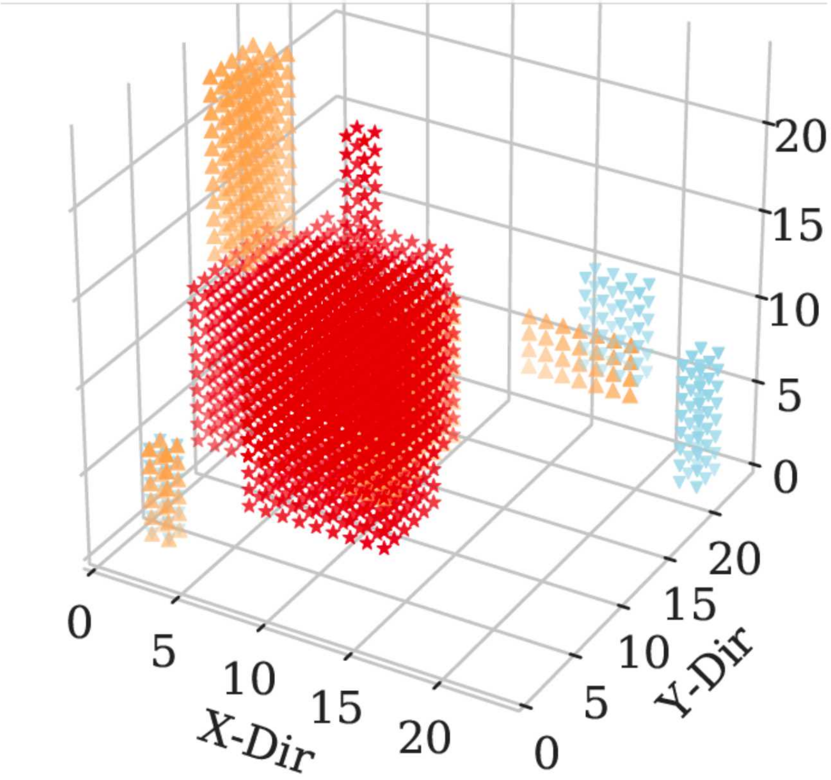
Visualization of Congestion in a Gemini Network



Plays at 10 real minutes per second

NOTE: Video Flattened in pdf

Automated Characterization of "Congestion Clouds" In Cray Gemini Networks



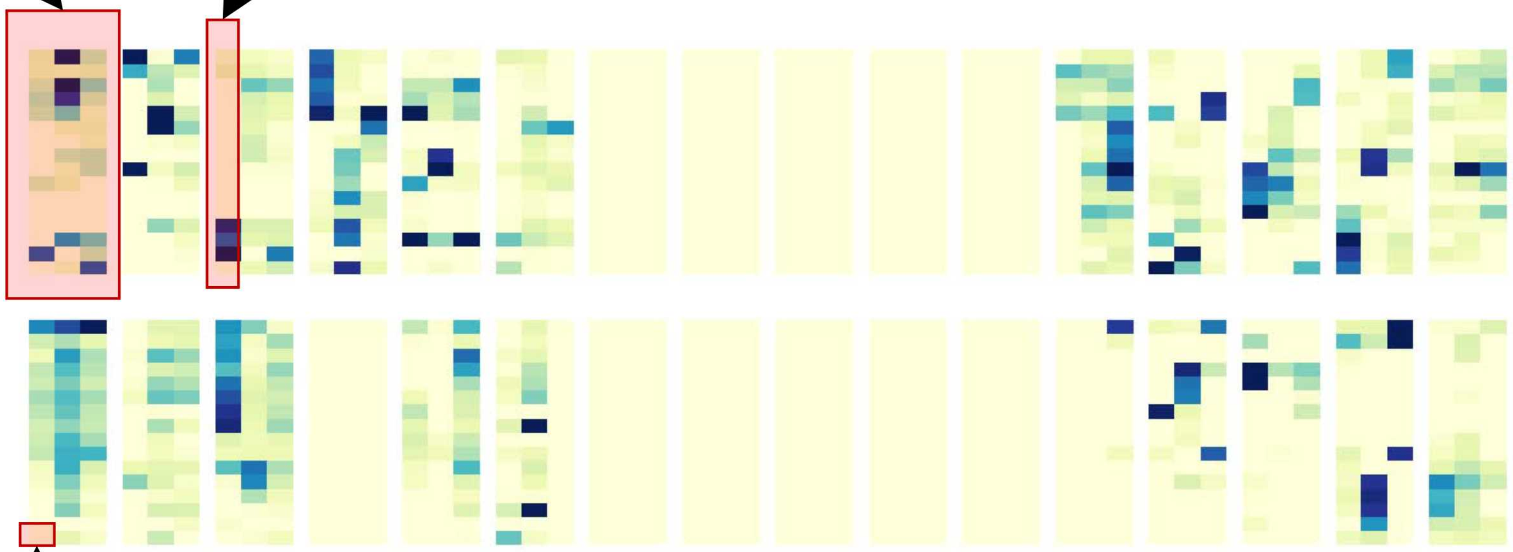
Long duration congestion clouds (direction independent)
(NCSA Blue Waters)

Analysis from NCSA's Blue Waters, *Lightweight Distributed Metric Service: A Scalable Infrastructure for Continuous Monitoring of Large Scale Computing Systems and Applications* SC14

Five Minutes of Congestion on Edison

Cabinet **Chassis**

NOTE: Video Flattened in pdf



>15% Percent Time Stalled (PTS) is dark blue
0% - 15% green gradient

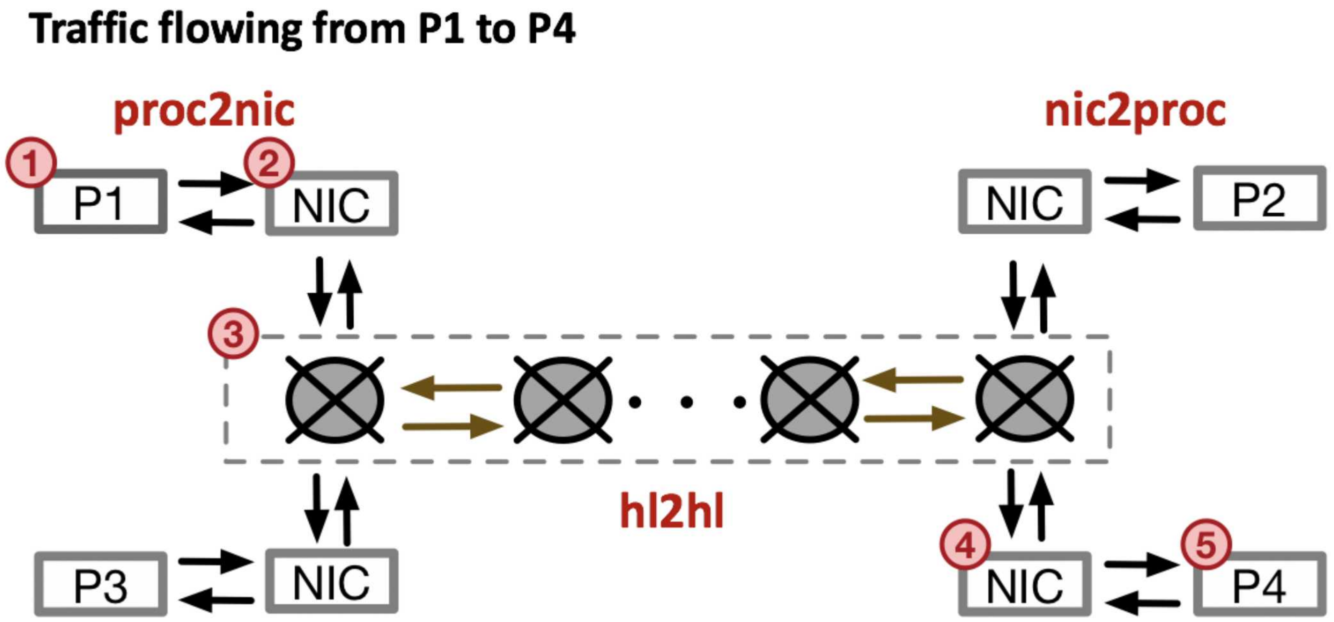
Aries Blade/Switch

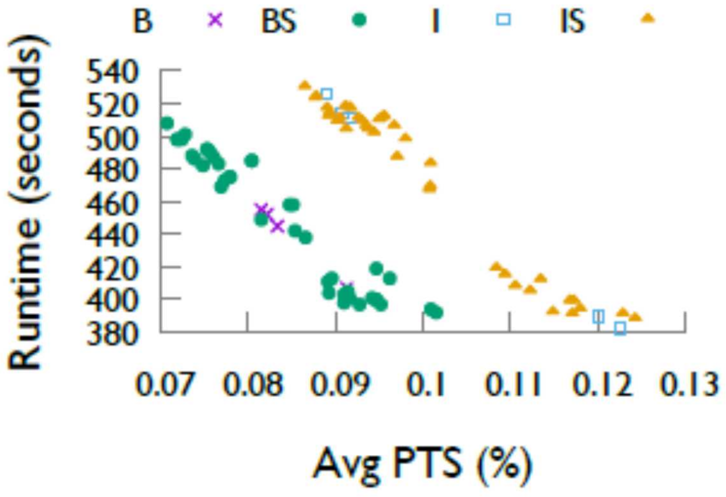
Network congestion assessments based on ~800 metrics per-router/sec
* 1000 routers in Edison

- Duration of congestion hotspots is short
- Hotspots move rapidly and application continues experience congestion

Indicators of Performance Impact

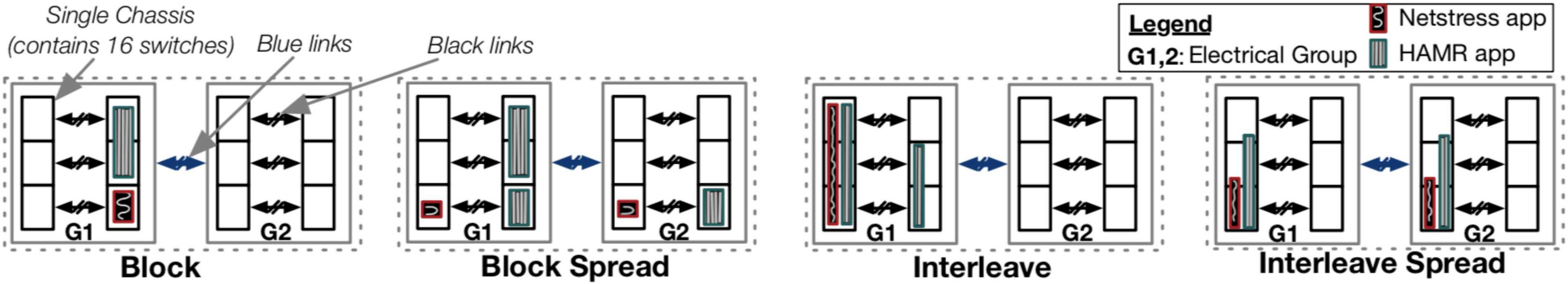
(allocation related)





(c) *nic2proc*

Illustrative Example: Scatter Plots of avg. PTS on link connecting NIC to processor and HAMR runtime

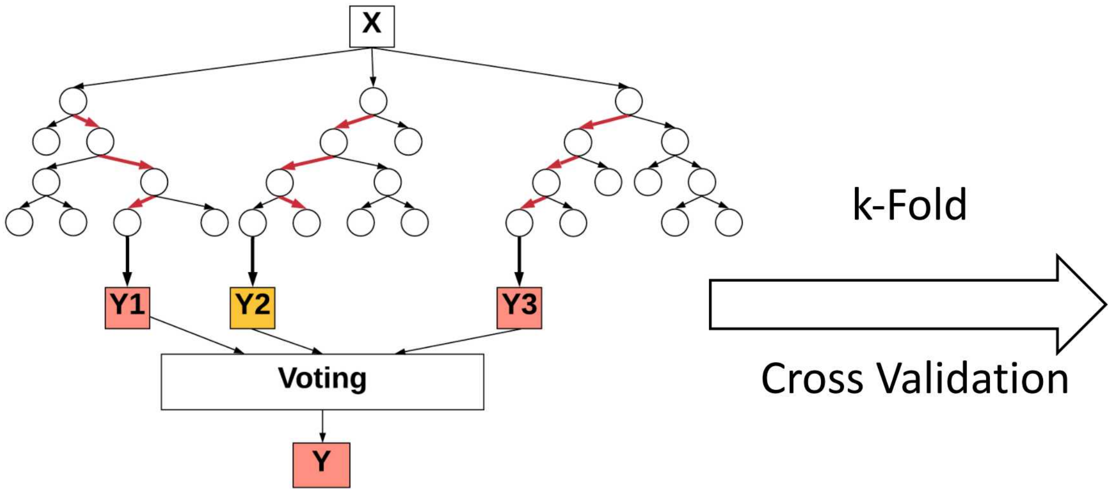


ML-driven Performance-Feature Discovery in Aries

Use of ML to identify Features most highly correlated with application performance

Information to be used for:

- Diagnosis for understanding performance variation
- Understanding performance sensitivity to each configuration parameter
- Scheduling policy



Ensemble Models

Feature Type	**Feature List**	**Avg R2***
Application and Scheduler Specific	*app,hugepages,placement, balance, avg hops*	0.64
Network Specific	*nic2proc,proc2nic,hl2hl* PTS & SPF	0.86
All features	All the above	0.91

*R2 is the coefficient of determination (higher is better)

Web Interface Example -- Showing Performance Figures of Merit for System and Application Performance Diagnosis

Job ID	App ID	Node ID	Runtime (s)	Back Pressure	Mem Score	Anomalies	PAPI Perf	App Perf
42093	miniAMR	nid000[52-55]	439	0.0	2	None	Back	1.45
42092	miniGhost	nid000[21;29-31]	1043	49.07	2	Cache	Back	-1.93
42091	miniMD	nid000[57-60]	617	5.24	3	Cache	Back	No data
42089	CoMD	nid000[52-55]	742	91.68	1	Cache	Back	1.52
42088	miniAMR	nid000[21;29-31]	447	0.0	2	Cache	Back	1.45
42087	miniGhost	nid000[57-60]	1043	73.88	2	Cache	Back	-0.27
42086	miniMD	nid000[21;29-31]	619	13.33	3	Cache	Back	No data
42084	CoMD	nid000[52-55]	742	90.88	1	Cache	Back	1.81
42034	miniGhost	nid000[52-55]	1022	98.59	1	None	Back	No data
42028	kripke	nid000[57-58	748	0.0	1	Mem	No data	No data
42027	kripke	nid000[21;29-31]	751	0.0	1	Mem	Back	No data
42019	kripke	nid000[52-55]	1092	0.0	1	Mem	Front, Back	No data

From: Enabling HPC Performance Insights via End-to-End Monitoring and Analysis – *in submission to IEEE Cluster 2019*

Mission Impact of SNL's HPC Monitoring Project

Sandia National Laboratories

- Developed, **deployed**, and continue to evolve, **efficient production-hardened HPC monitoring software stack**
- Widespread use is enabling:
  - More data --> more measurement based understanding and better HPC platforms
    - Across DOE complex – Tri-Labs, NERSC, NCSA
    - Cray
    - Other sites using HPC resources
  - User **contributions** that provide additional functionality at **no additional cost** to DOE
- **Cray's adoption** means integration into DOE's next generation mission computing platforms from the outset
  - Collection of data during high occurrence of rare event period
  - Researchers can spend their time on data analysis and feedback aspects → more efficient computing both now and in the future

Planned & Ongoing Work (next 3 – 5 years)

- LDMS:
  - Enhancements to core and plugin capabilities
    - Ease of configuration
    - Expanded and enhanced transports
    - Plugins to support new architectures, subsystems, and functionalities
  - Direct support for efficient scalable storage architectures
  - Continue to improve on overhead and robustness to system failures
  - Incorporate new data transport constructs to keep in step with changing technologies (e.g., pub/sub technologies, json formats)
- Advanced analytics:
  - Application and development of analytics with domain-relevancy
  - Integration of log and facilities data
  - Explainable machine-learning methods for extracting responses and future designs
    - Assessment of confidence in analytical results and associated response options
  - Research and development of technologies to facilitate higher efficiency storage and queries to support advanced run-time analytics
  - Continued collaborative research, including experiments and analysis of production scenarios on extreme-scale systems of collaborative partners
- Feedback
  - Automate Feedback of actionable data to applications and system software (e.g., schedulers)
  - Enhanced forms of feedback to users and operations staff
- Production Sustainability
  - As we develop, validate, and production harden components of both analytics and monitoring infrastructure we will continue to deploy both at SNL and at partner sites
  - Building a self-sustaining support community – currently have a users group, tutorials, documented contribution processes

Questions?

- Project Objective
  - **Increase** the **performance** of current and future HPC systems through improved, **data-driven approaches** to operations and systems design.
    - Improve understanding of application ←→ system interaction
    - Automate run-time data-driven responses to problem identification
- Strategy/Goals
  - Perform continuous high fidelity **system wide monitoring** on large scale HPC systems (SNL and collaborator sites)
  - Gain **new insights** about application and HPC sub-system interactions **through large-scale analysis** of whole-HPC-ecosystem data
  - Utilize new understanding to build platform-independent infrastructure and tools to **enable effective automated run-time analysis and feedback**
- Technical Challenges
  - Whole-system data **collection** and transport **cannot impact performance**
  - **Data management** must efficiently support run-time and historical **analysis**
  - Extraction of actionable information (**analysis**) from **high volume** (~10s of TB/day) and **high dimension** (100s to 1000s of discreet variables) data
  - **Relevant** raw **data** may be **un-exposed or unknown** and change from architecture-to-architecture
  - How raw and derived **data** can be **quantifiably associated with application impact** may be unknown
  - Explainable and interpretable Machine Learning (ML) application to **active feedback** systems
  - Analytical approaches need to support incorporation of **architecture-relevant information**, and provide **confidence bounds** for response
  - **Unknown** which **application** runs should have similar performance, to detect abnormal performance
  - Getting **early access** to machines: “rare” events occur more frequently
  - Cleansing and **releasing data** for analysis research is problematic