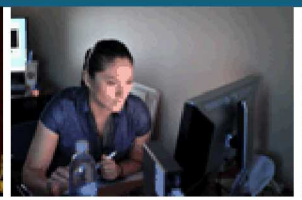
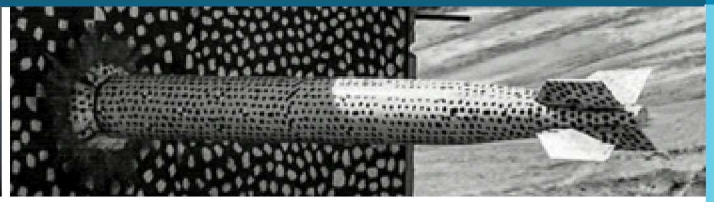
SAND2019-8090PE

Lightweight Distributed Metric Service: Deployments, Enhancements, Roadmap, and Activities





Jim Brandt (brandt@sandia.gov)

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & 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.

Acknowledgements

Significant contributions from:

- Sandia National Laboratories (SNL)
- Open Grid Computing (OGC)
- Los Alamos National Laboratories (LANL)
- National Energy Research Scientific Computing Center (NERSC)
- National Center for Supercomputing Applications (NCSA)
- University of Illinois at Urbana-Champaign (UIUC)
- New Mexico State University (NMSU)
- Boston University (BU)

Outline

- Brief LDMS Overview
- Overview of LDMS v4 & v5
- Release Timeline
- Utilization of LDMS Collected Data For Analysis
- Deployments at SNL
- Getting Involved

Lightweight Distributed Metric Service (LDMS)

- Synchronized system wide data sampling provides resource utilization "snapshots"
  - Memory, Memory Bandwidth, CPU, Power, Hardware performance counters
  - Network utilization and congestion parameters, I/O
- No significant impact on applications at collection rates (1Hz or higher) necessary for resolving resource utilization features
  - **Optimized data structures**
  - **Optimized Transport** over RDMA, IB, socket
    - The RDMA transport enables aggregation and storage of sampled data at no additional compute node CPU cost
  - Testing at scale on Blue Waters (27,648 nodes) and Trinity (2 * 10,000 nodes)
    - ~ 5TB/day on Trinity
  - On-node CPU overhead limited to the cost for data exposure and a single memory copy

Scalable and low overhead acquisition, transport, and storage of information

Features In v4

Security: More Control Over Set Access

- Plugin based authentication now supports none, shared secret, and munge based authentication for access to ldmsd data and meta-data

```
voltrino:~ # ldms_ls -h localhost -x sock -p 411 -a munge
nid00006/vmstat
nid00006/var/opt/cray/imps/image_roots/login-large_cle_6.0up05_sles_12sp3_x86-64_ari-created20180529.SNL
nid00006/var/opt/cray/imps
nid00006/procstat
nid00006/meminfo
nid00006/jobinfo
nid00006/cray_aries_r_sampler
nid00006/aries_rtr_mmr
nid00006/aries_nic_mmr
nid00006/aries_linkstatus
```

```
gentile@voltrino:~> ldms_ls -h localhost
nid00006/vmstat
nid00006/jobinfo
```

Metric sets seen based on permissions via munge: Root (top) sees more than user (bottom)

```
voltrino:~ # ldms_ls -h localhost -x sock -p 411 -a munge -v nid00006/meminfo
nid00006/meminfo: consistent, last update: Mon Jul 02 11:07:09 2018 [2540us]
  APPLICATION SET INFORMATION ------
              updt_hint_us : 1000000:0
  METADATA --------
    Producer Name : nid00006
    Instance Name : nid00006/meminfo
      Schema Name : meminfo_x86_ven0000fam0006mod002D
             Size : 1976
     Metric Count : 46
               GN : 2
             User : root(0)
            Group : 44476
      Permissions : -rwxrwx---
  DATA ------------
        Timestamp : Mon Jul 02 11:07:09 2018 [2540us]
         Duration : [0.000035s]
       Consistent : TRUE
             Size : 416
               GN : 16923
  ----------------
```

Update hint

Permissions

Ease of Set-up and Dynamic Reconfiguration

```
voltrino:~ # ldms_ls -h localhost -x sock -p 411 -a munge -v nid00006/meminfo
nid00006/meminfo: consistent, last update: Mon Jul 02 11:07:09 2018 [2540us]
  APPLICATION SET INFORMATION ------
              updt_hint_us : 1000000:0
  METADATA --------
    Producer Name : nid00006
    Instance Name : nid00006/meminfo
      Schema Name : meminfo_x86_ven0000fam0006mod002D
             Size : 1976
     Metric Count : 46
               GN : 2
             User : root(0)
            Group : 44476
      Permissions : -rwxrwx---
  DATA ------------
        Timestamp : Mon Jul 02 11:07:09 2018 [2540us]
         Duration : [0.000035s]
       Consistent : TRUE
             Size : 416
               GN : 16923
  ----------------
```

Update hint

Permissions

- Aggregator can automatically adjust to collection rate changes through update hint in the metadata.
- Easier to adjust collection rates during events of interest

More Control Over Set Collection and Aggregation

- Vector of sets - A sampler can collect, and locally store, multiple time instances of the same metric sets (e.g., for *high frequency collection*).
  - *Sets can be aggregated at lower frequency than collection, without loss of data.*
  - *Can be used to increase reliability of data gathering in the face of short network outages*
- Metric Set Groups - Enables defining groups of metric sets that may come and go and only require a group update to be defined for all sets within the group
- Automated Failover - User only needs to define pairwise sets of aggregators responsible for picking up the aggregation load if one or the other should fail
- Push/Pull Over Both Sock and RDMA - pair-wise configuration between two ldmsd with a producer/consumer relationship
- Support for multiple transport/port configuration - *A single ldmsd can listen on multiple transports and/or ports simultaneously*

Job Association Capability

- SLURM Job Id Sampler - Adds Job Id into all collected sets for ease of per-job analysis

v4.3.1 and Greater

- Set plugin destroy removes all plugin related data structures
- Note that v4.3.x is transport incompatible with v4.2.x and earlier

Features In v5

Expanded Plugin Capabilities

- Expands sampler plugin functionality and supports multiple instances, with different configurations, of the same plugin on the same ldmsd (e.g., multiple Lustre sampler instances each sampling for a different mount point)
- Expand plugin types
- Per-plugin state and status information
- Configuration driven plugin<—>plugin interaction

Consolidated Configuration

- New sampler configuration object to normalize configuration of sampler to that of other configuration objects
- Configuration redesign that enables complete configuration using a single file rather than combinations of files and environmental variables

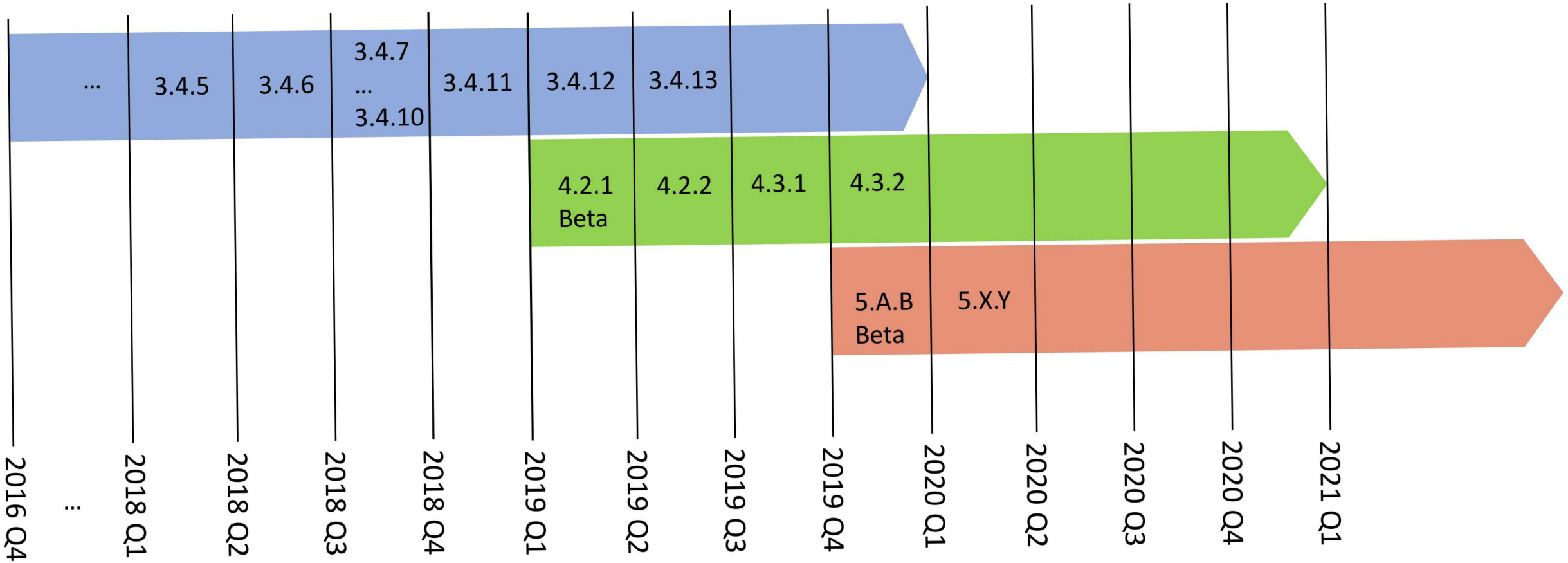
Expanded Transport Capabilities

- Support for RDMA over libfabric
  - Support Intel OPA and Cray Slingshot

Miscellaneous

- Support for pub/sub over the LDMS transport
  - Like Rabbit but with RDMA support
- Metric sets with the same schema can now have a variable number of elements
- Storage plugin to support the Distributed Scalable Object Store (DSOS)
  - DSOS being released at the same time

Release Timeline

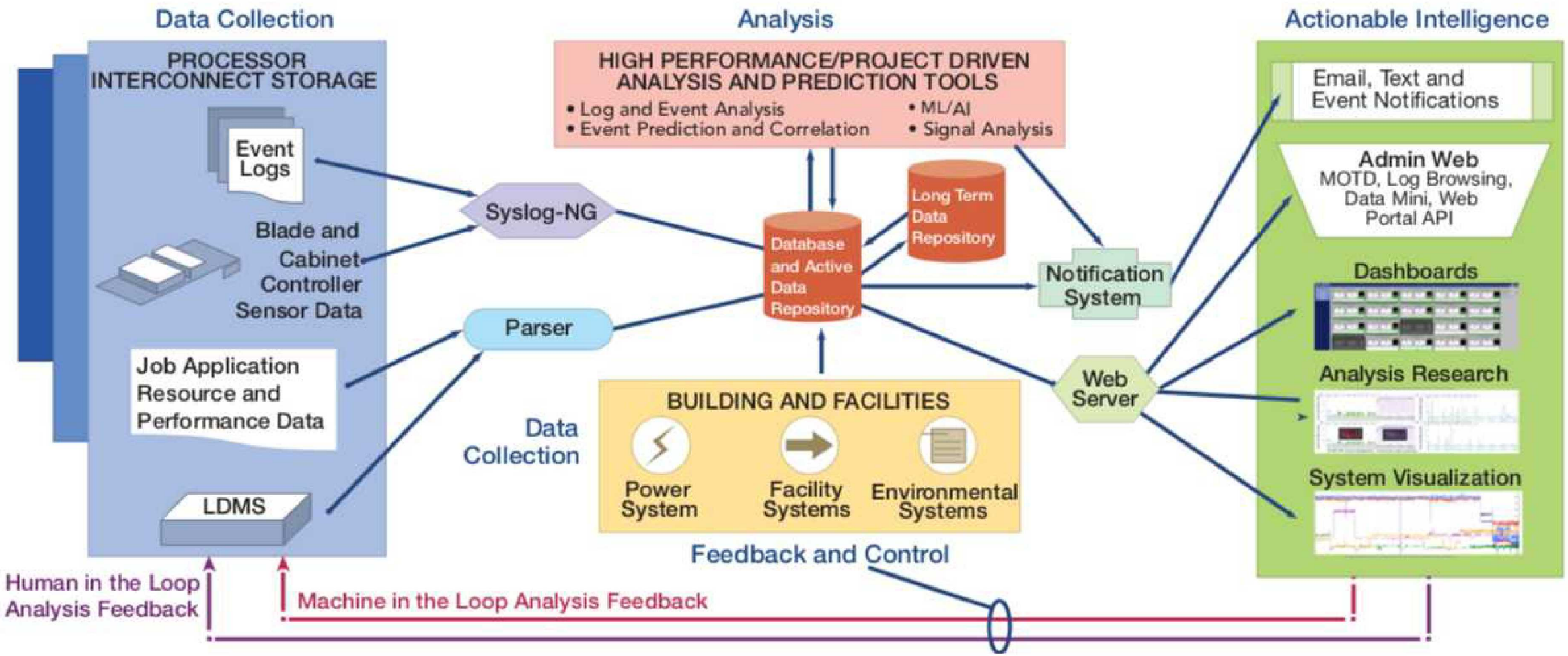


Utilization of LDMS-Collected Data For Analysis

Key Questions

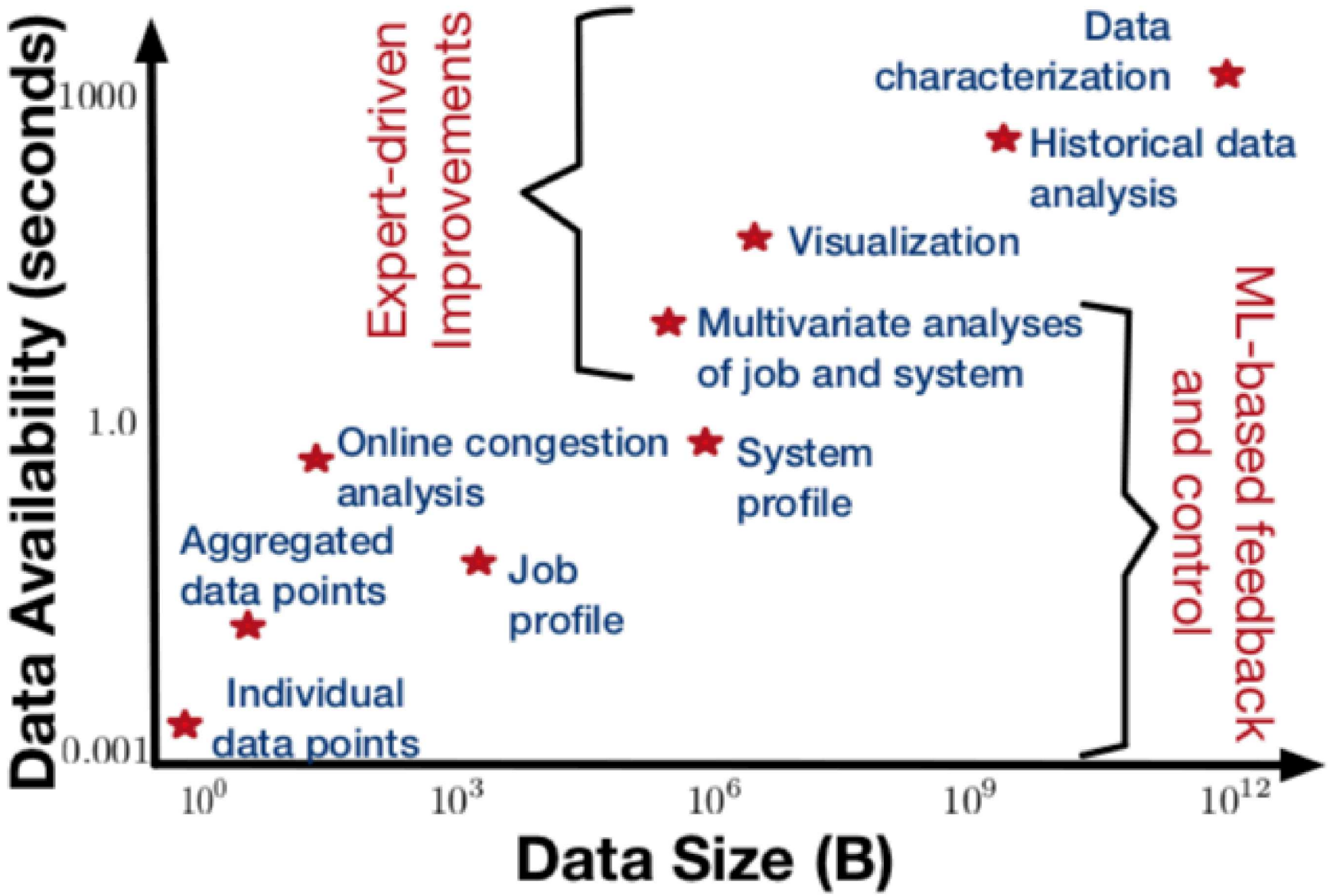
- System Managers and Users: How can I know if the system is having problems?
- Users: Is **my** application performance variation due to system conditions or code changes?
- Architects and Acquisition Teams: What are the architectural requirements given the site's workload?
- Architects, System Managers, and Support Staff: How can a system provide more effective and efficient services?

Holistic Measurement Driven System Assessment (HMDSA) Scalable System Architecture



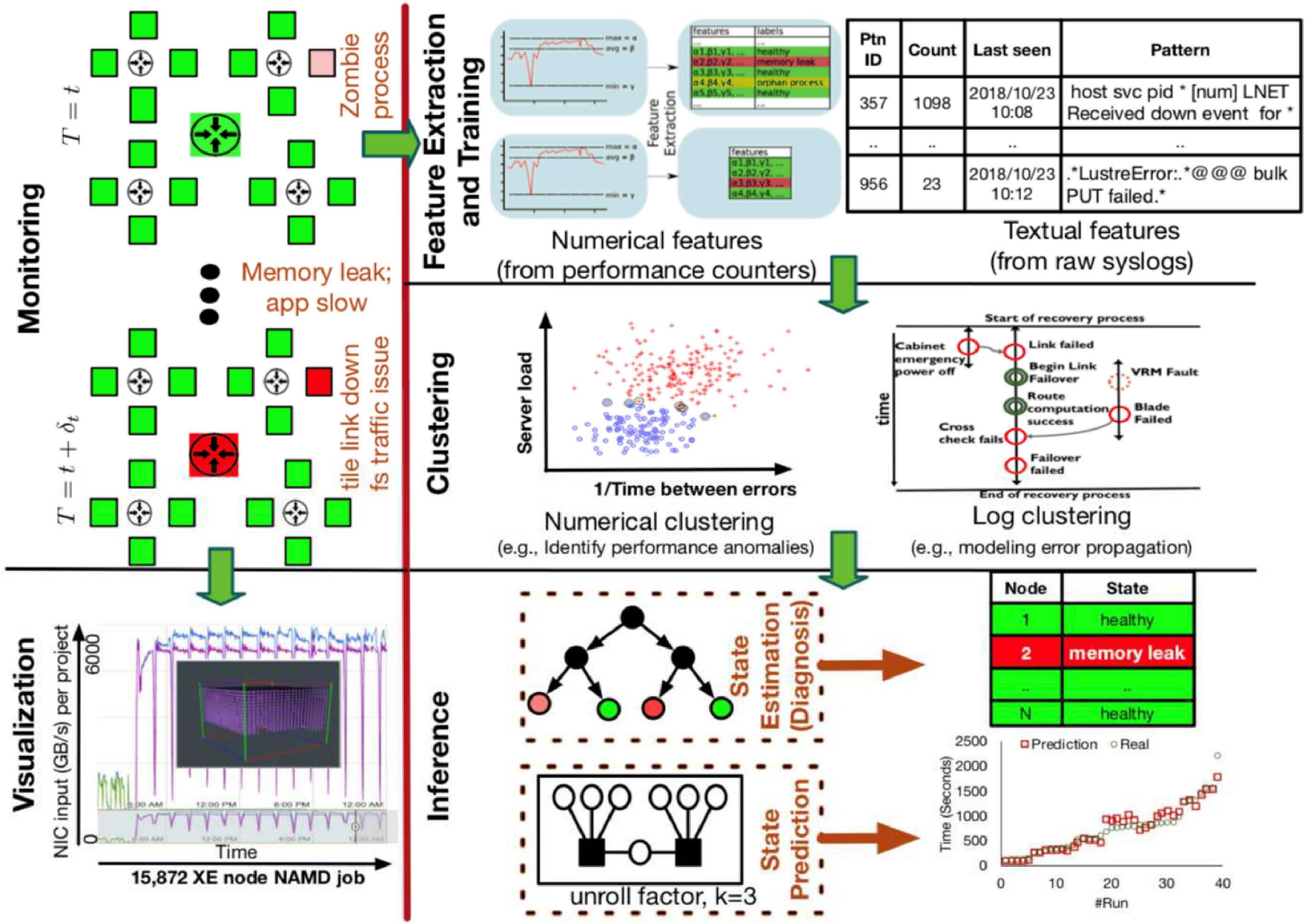
From: Holistic Measurement Driven System Assessment – *ECP Annual Meeting 2019*

Actionable Time Scales



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

Actionable Expert- and ML-Driven Analysis



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

Analytics Results Displayed on E2EMon Interface

Edit profile

Profile: fred Submit

Pane Settings:

- Enable pane "Cluster" at position: 1
- Enable pane "Center" at position: 0
- Enable pane "User" at position: 2
- Enable pane "Queues" at position: 4
- Enable pane "Jobs" at position: 3
- Enable pane "Job" at position: 5
- Enable pane "LDMS" at position: 5
- Enable pane "Explore" at position: 6

Report Settings:

- Enable "computes_cluster" on pane: Cluster at rank: 0
- Enable "work_mix_cluster" on pane: Cluster at rank: 1
- Enable "hsn_cluster" on pane: Cluster at rank: 1
- Enable "hardware_cluster" on pane: Cluster at rank: 1
- Enable "queue_summary_cluster" on pane: Queues at rank: 1
- Enable "computes" on pane: Center at rank: 0
- Enable "work_mix" on pane: Center at rank: 1
- Enable "hsn" on pane: Center at rank: 2
- Enable "hardware" on pane: Center at rank: 3
- Enable "nfs" on pane: Center at rank: 4

OVIS for CAPVIZ: Jobs - Mozilla Firefox

OVIS for CAPVIZ: Jobs

http://localhost:8080/capviz/testuser/Jobs/

User: testuser Profile: testuser

Center Cluster User Jobs Queues Job LDMS

[?]

Voltrino Job Information

From:

To:

Show 25 entries Search:

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
42090	kripke	nid000[21,29-31]	66	0.0	1	None	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
42085	kripke	nid000[57-60]	66	0.0	1	None	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
42030	kripke	nid000[21,29-31]	709	0.0	1	Mem	Back	No data

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

Job Based Drill Down

E2EMon: Job Detail

back to job listing

Job ID	App ID	Node ID	Runtime (s)
42093	miniAMR	nid000[52-55]	439

Back Pressure	Mem Score	Anomalies	PAPI Perf	App Perf
0.0	2	None	Back	1.45

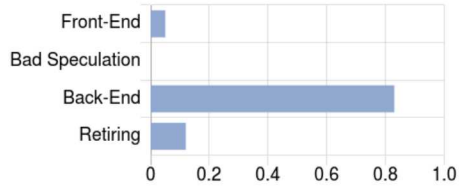
Application Heartbeat Analysis:

Application heartbeat data for job 42093, [0, 1, 2, 3, 4] (graph)

HBeat #	Avg	Dev	Min	Max	MinPID	MaxPID	Graph
0	4.41	1.45	4.28	4.64	55442	55856	graph
1	88.26	28.97	85.69	92.88	55442	55856	graph
2	0.0	0.0	0.0	0.0	24050	24050	graph
3	3530.3	1158.65	3427.57	3715.3	55442	55856	graph
4	0.0	0.0	0.0	0.0	24050	24050	graph

TopDown Analysis:

First Level TopDown Analysis for job 42093



Threshold for reporting Retiring is .75, threshold for all other bottlenecks is .1

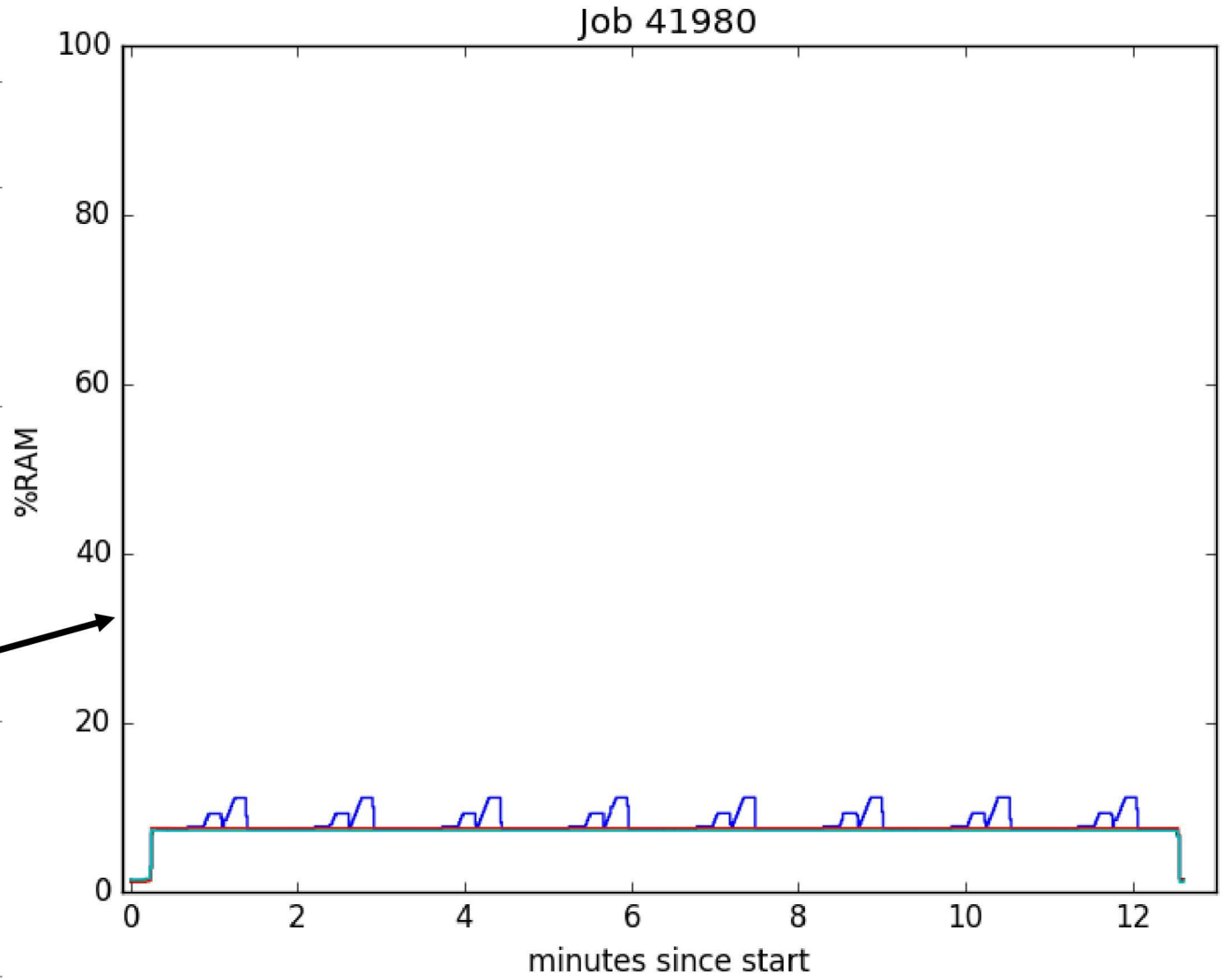
Memory Analysis for 42093:

Avg %RAM	Stdev %RAM	Max %RAM	Min %RAM peak
9.8	1.6	10.4	10.3
Sample interval	Balance	max %RAM host	Min %RAM peak host
1	1	52	35

VIEW	Filesystem location
histogram	/lscratch/e2emon/data/memplot/4209303000000000.png
line plot	/lscratch/e2emon/data/memplot/4209302000000000.png
job end UNIX UTC	1558735473

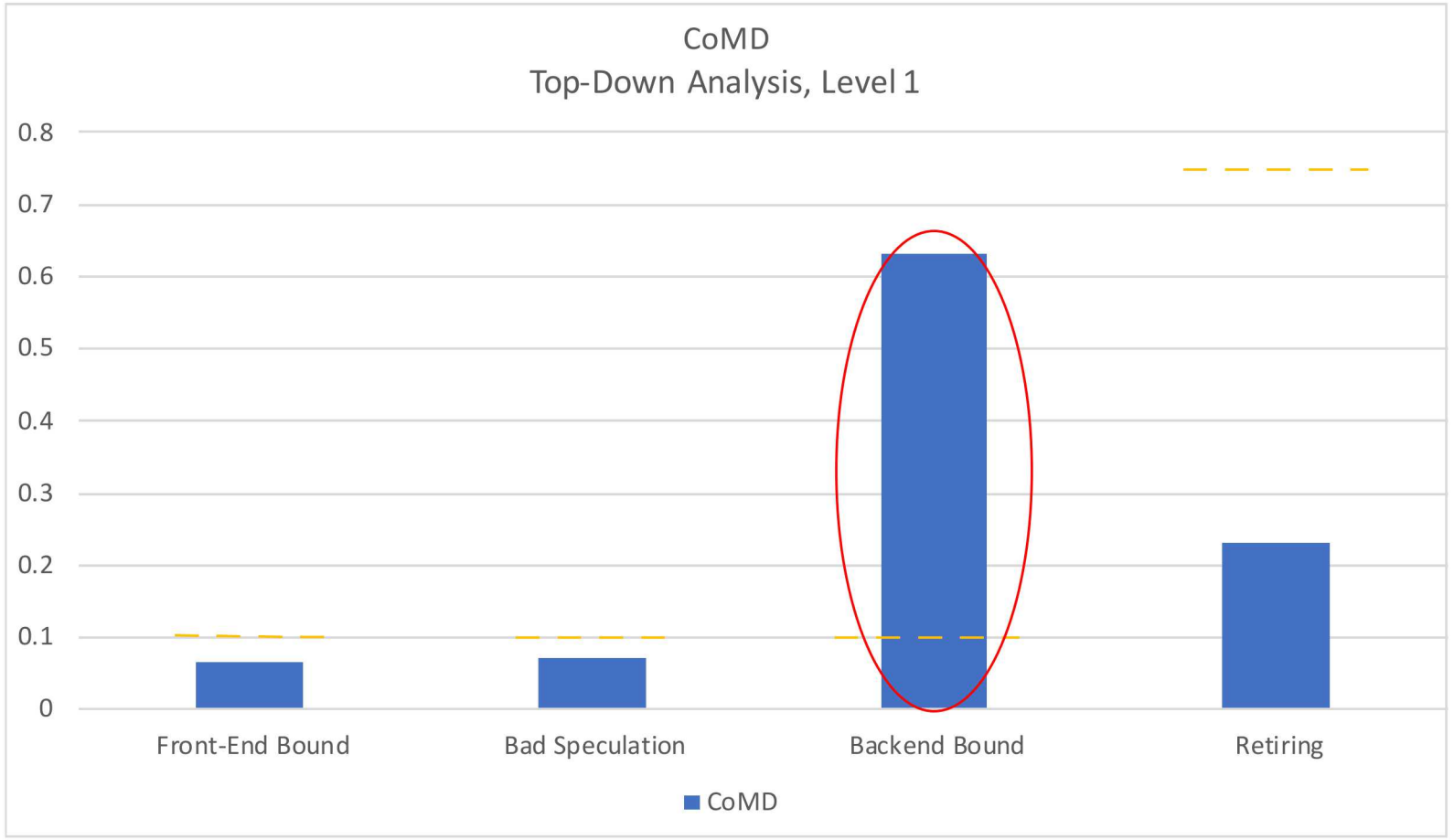
%RAM is computed as 100*(MemTotal-MemFree)/MemTotal

No detailed network data available



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

Top Down Analysis (TDA)
Based on Intel Top Down Analysis Method



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

E2EMon Interface Example Use Case

Same Inputs Different Behaviors

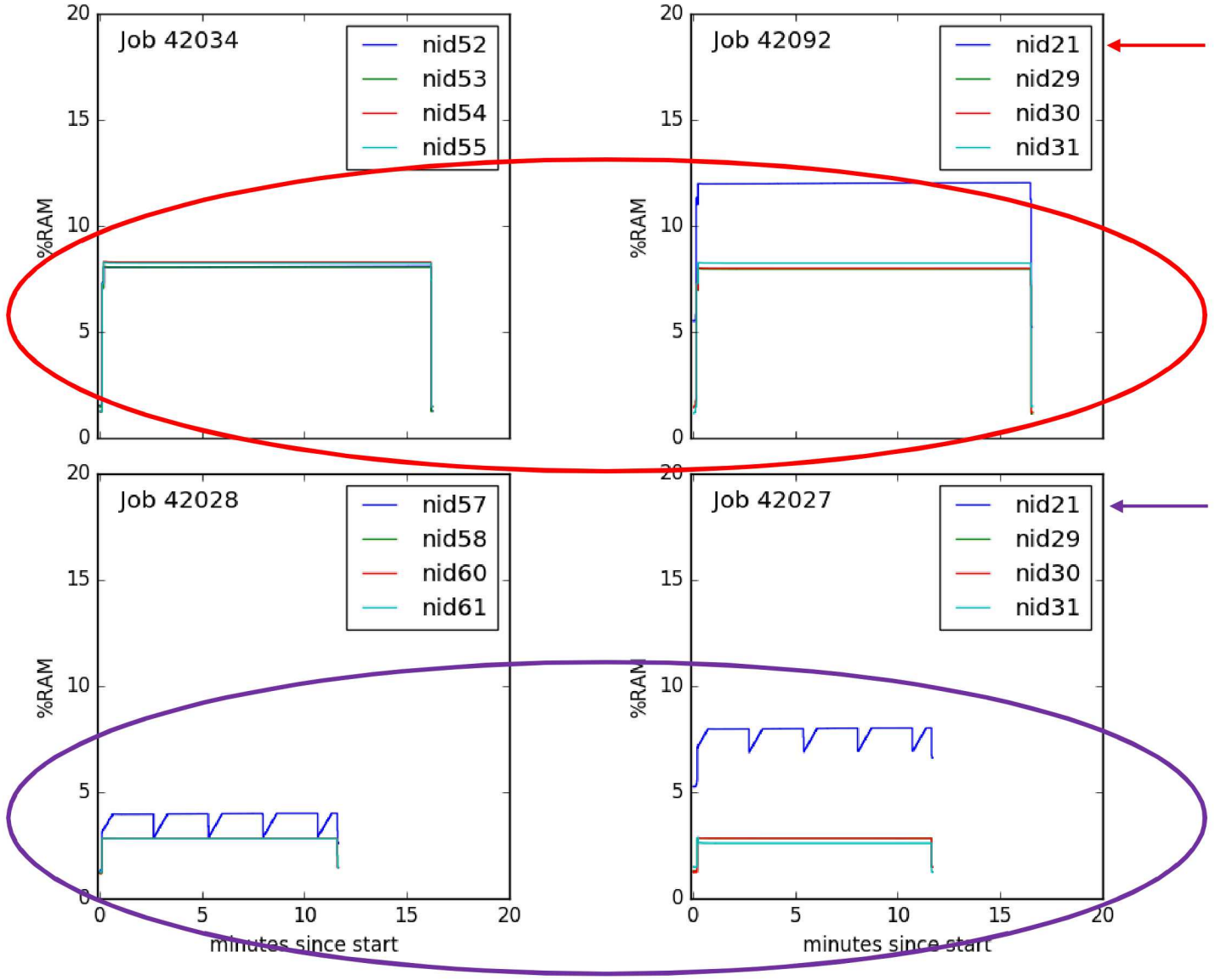
Job ID	App ID	Node ID	Runtime (s)	Back Pressure	Mem Score	Anomalies	PAPI Perf	App Perf
42093	miniAMR	nid000[52-55]	439	0.0	2	None	Back	1.45
42092	miniGhost	nid000[21;29-31]	1043	49.07	2	Cache	Back	-1.93
42091	miniMD	nid000[57-60]	617	5.24	3	Cache	Back	No data
42089	CoMD	nid000[52-55]	742	91.68	1	Cache	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*

Same Inputs Different Behaviors Drill Down

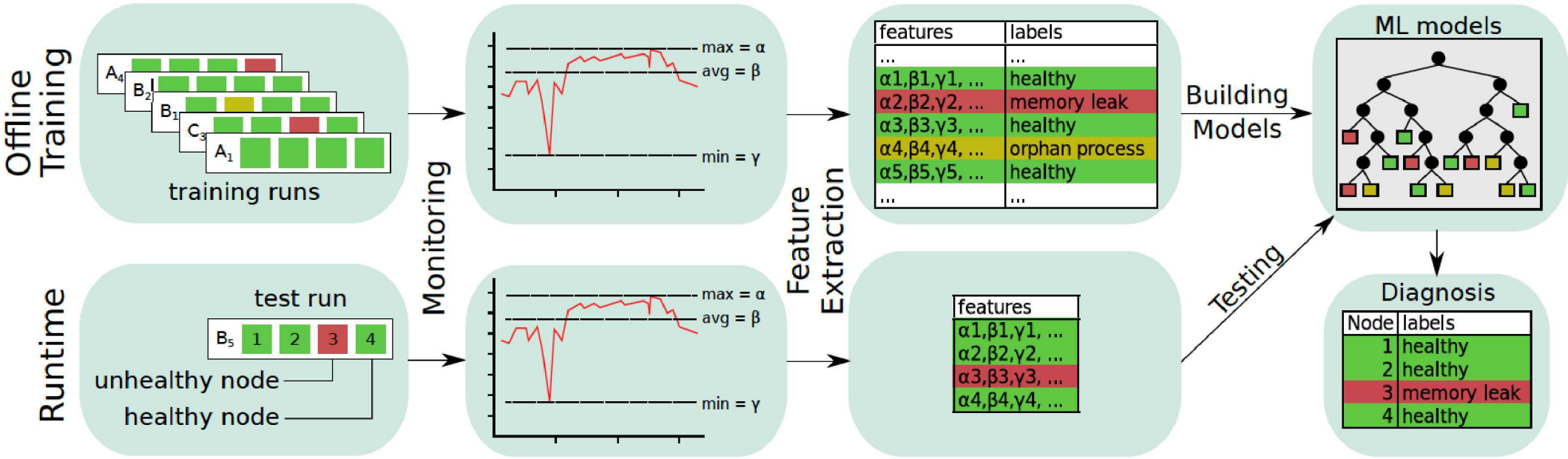
Two identical runs of miniGhost

Two identical runs of kripke



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

Anomaly Detection and Problem Diagnosis



Detection and diagnosis of performance problems

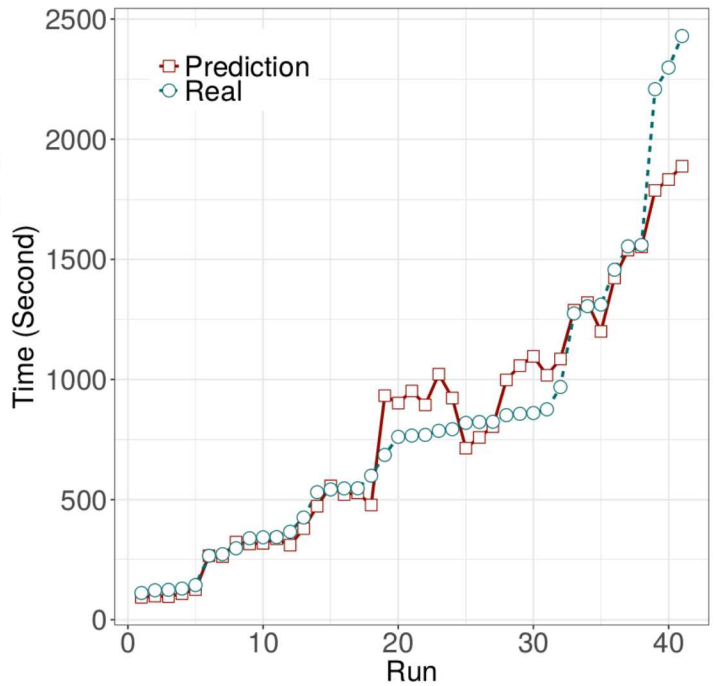
- Machine learning models built offline are used for classifying observations at runtime
- Detect and diagnose behavioral differences due to: memory leaks, errant processes, contention, etc...

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

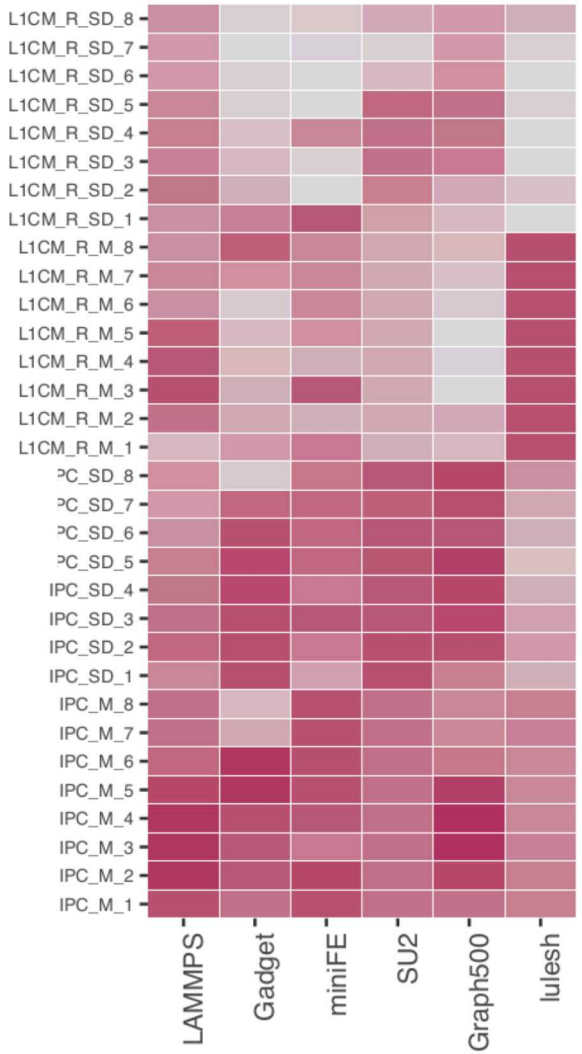
Application Progress Assessment and Performance Prediction

- User-determined data collection and analysis based on specific application details
- Analysis: Use application data to assess application progress and sensitivity
- Figure of merit metric: Avg. heartbeat1 rate

Detect and predict performance problems based on variation of run time progress in sensitive code sections



Interval h/w counter importance heatmap



LAMMPS runtime prediction

LDMS Deployments on Production and Testbed Systems at SNL

- Deployments on Production Capacity systems (v3.4.13)
  - All but one production systems (9,553 hosts total) are running LDMS
  - Sampling intervals are all 60 seconds
  - Data currently largely being utilized to analyze network performance/behaviors
- Deployments on Testbed systems (v4.2.3)
  - Mutrino and Voltrino – Trinity Application Readiness Testbed (ART) system
    - Sampling intervals: 60s (link status), 100ms (power & energy), 1s other
  - Other Testbed systems
    - Currently running on 6 of 12 systems
    - Sampling intervals of 1s for all samplers
  - Data being used to enable HPC analytics and feedback research as well as application resource utilization and performance understanding
- In-situ deployments are also utilized to enable users access to “canned” metrics of interest including a variety of Hardware Performance Counters

Getting Involved

- Discussions at: https://github.com/ovis-hpc/ovis/issues
- Participate in LDMS Users Group bi-weekly zoom meeting
  - Telecon info at: https://github.com/ovis-hpc/ovis/wiki/User-Group-Meeting-Notes
- Get help
  - Quick start build and configuration information at: https://github.com/ovis-hpc/ovis/wiki under “LDMS v4 Documentation” sidebar
  - Post problems to: https://github.com/ovis-hpc/ovis/issues
- OVIS/LDMS Users Group 2-3 day Face to Face meeting in October 2019 at University of Central Florida campus in Orlando Florida
  - Will post to wiki as plans solidify: https://github.com/ovis-hpc/ovis/wiki

Questions?