Website and mailing list: ***https://sites.google.com/site/monitoringlargescalehpcsystems/***
Signup for starting discussion in a topic area or writing a quickstart guide: HERE

How to keep momentum on a topic?

- **Webinar? Every other month?**
- **Quick start guides**
  - **Help get something stood up in a short amount of time**
  - **Can more quickly convince funding sources that something is a viable path**
  - **Seed Quick start guides with a scenario. Then the guide can describe how you would address the scenario with that tool**
  - **There is interest in writing and using these**

Interests:

- **Survey in the general state of the art of HPC monitoring**

1) Analysis capabilities (pandas, spark)

There are a lot of new analysis capabilities, like the SciPy tools, Spark, etc. that I would like more info on:

- When are these useful?
    - What features? What types of data? What types of analyses?
- Do I need to have a lot of data or distributed data stores?
- How hard are they to use or write new analysis?

Spark - not many
SciPy - not many
R
MPI job
Using facilities/visualization within monitoring tools - ganglia, nagios, slurm.

- Concerns with some of these tools for long term analysis

Graphana, open xdmod

If you aren't doing analysis, why not?

- Don't know what to look for

Why do you want to do analysis?

- Operations needs (immediate)
- Historical comparisons

- Finding abnormalities
- Research
    - Limited by access controls
    - Cleanser tools?

How are you doing analysis:

- Post processing
- As needed as issues arise / Realtime
- Linking jobs (nodes) to events.
    - Sub-node allocations?
    - Ad hoc tools for making these associations at the sub-node level

Research that is wanted:

- Fingerprinting workloads and using those for queueing decisions, associating with errors etc.

Publically releasable dataset and systems:

- SNL is putting up a machine for analysis. Would have data sets that researchers could try out their analysis tools w/o having to move the data.
- USENIX site with logs

Job data, system data, performance counters?

Making state data available to users so they know where issues/bottlenecks/load etc is

2) Application Impact

I want to know about analyses and data to assess application impact due to system events, system load, and interfering applications:

- What types of data should I collect?
- How can I determine if an application has been impacted given variable production load? And by how much?
- How do I do attribution of the source of events?

Proving monitoring doesn't impact job performance?

- Don't ask :)
- Have to be cognizant that there is a range of overheads for various tools
- Design of experiments:
    - including topological placement
    - consistent run time at scale
    - Other system services
    - Shared filesystems with other systems at the same site
    - Application choices - mem BW, interconnect latency, interconnect BW, computationally intensive
- Would like resources on measuring impact and design of experiments, application features

- Working testing into your normal production loads?
    - Do you have variable workloads over time, so that throughput variation is or is not a valid metric?

Monitoring rates to get useful info vs overhead

-

Ensuring validity as you change up your monitoring system and system software?

Monitoring features to support analysis?

- Synchronous collection supports comparing data values across nodes in job/system

Is monitoring part of the agreement?

- Some include that they can use the data in anonymized form

Is behavior part of the agreement?

3) I/O

I want to know about existing tools/techniques for getting information on I/O, including:

- Do I need more system resources to satisfy the demand?
- How can I know what applications are hammering the system and how they can change what they are doing to improve performance?

Monitoring user behavior

How can I improve my IO throughput via measurement and resources

Do people measure IO? Using what tools?

- Instrument user codes via DARSHAN (wrappers for code). Papi counters.
- Instrument MDS, server & client side, parallel FS. watch I/O ops per sec. Schedule based on historical user/app behavior data.
- Lustre + PBS - utilization data (Lustre Stats) (R/W/open/close, per job etc).
- Pcp data for I/O

Want better understanding of the workload demands

- For scheduling
- For resource requirements
- Feedback to users

How many people find it easy to id apps abusing the systems?

- Variable workloads -- can't use historical run behaviors
- (A few) Open/close, which jobs etc. is easy. If it's 1 job, then easy. We don't have good idea of combinations - this one doing seeks, that one doing opens etc. Can do node associations and don't have multiple jobs per node.

User Behaviors

- Users Creating too many files? (many)
- Non-ideal use of the resources - many small files? Write sizes
- Use quotas initially to prevent problems. Then they get more space if they are important.
- Live testing to determine system status/response

Lustre - performance characterizations changes.

Vendors not understanding why you want the info you want? Don't give you the info you want?

- We don't want a new GUI with another subset of the data. Instead give us a way to get the data.

4) Network contention

I want to know about existing tools/techniques for understanding:

- Is there contention in the network?
- Are network conditions (contention, available bandwidth) impacting application performance?
- What applications are responsible for the contention?

How many people want to know this?

- A few
- No one says they dont have network problems

What tools are you using to get the data and analyze it

- IB - vendor specific interfaces (Mellanox) or OFED command line tools
- Don't know about getting info from links between two switch chips. The info on a link between a switch and a node is easier for us to gather and understand.. Or w/o causing side effects
- Indirect measures
- Error counters

Configurations, application configurations to reduce network congestion

SGI MPT

Omnipath - tools? Performance?

Upcoming switches and instrumentation?

System network

- Congestion events, locations, sphere of impact

Network into the data center

- Users complain about not being able to transfer wide area data. Need to ID where the issues are. Publish the network state data on the web so users see where the problem is (or is not)

How frequent?
What do we monitor?
What are side effects of monitoring?

Overlay networks for collecting data and reducing data

- **Scalable overlay network (scon)**
-

5) Data stores

- I want to know about options for storing text and numeric log data, including:
    - Performance for handling large amounts of data
    - When should I worry about in-memory vs. on-disk?
    - What tools exist that work with the stored for visualization and analysis?
    - How easy/hard to set up? Best practices for configuration

Data aggregation:

- Aggregating from multiple machines?

SQL
Maria
INfluxDB
OpenTSDB
ELK stack
Cassandra
Ceph (object store)
Graphite
Redis

Identify data model, analyses, ingestion rates

What stores handle what kind of data/queries?

- Xdmod - uses fairly standard store but setup for its analysis

Data rates?

Supporting analysis vs viz vs ingest rates

Unhappy with the storage solution you have? What is lacking?

- Collecting using adhoc tools (script) and writing to a file, but I don't have time to write better
- Dumping script data in a database/tool to present a better face on it
- Thrown away huge amounts of data because we couldn't store it
- Have a working infrastructure that we can't re-engineer due to time

Quick check vs production level continuous monitoring

6) Visualization and dashboard tools

- I want to know about tools that exist with which I can use to easily make dashboards:
    - How much coding is involved to use them?
    - Are there access control features?
    - Can I share a dashboard I created with someone else?
    - What kinds of plotting does it support?
    - What is the performance for large number of nodes or long periods of time?

- Alerting - false positives, cascading failures
- Tools - Ganglia, Nagios, Graphana, pcp (part of pcp is pmie), feed pcp data into Nagios
    - Many people use G & N for their visual dashboards. May not be using in subsequent analysis.
- What can we learn for tools, visualizations, analyses from other domains?
    - Tradeoffs of performance for these tools for other non-HPC domains
- Web interfaces (php)
    - Does any info need access controls? Yes, but this is typically either done as admin only or all access

    - How to share system data with researchers (Access control)?
- Desires of what to see:
    - Realtime heat maps for users to see what jobs can get in
    - Users wanting to see their job data?
- Many people are looking for other visualization tools
    - Using graphana now. Good for 1st pass. Want more sophisticated templating. Use with a database. Don't want a monolithic thing. Want to do subsets of components. SOmeone else backs that with influx DB instead of an SQL db.
    - D3, google charts - can front a database with a php query that can then spit the output into google graphs (people would like a quick start guide on that)
    - XDmod/Supreme. Plugin that can break it down at a job level.
    - Ganglia Jobmonarch - marry slurm with ganglia
    - Splunk - cost prohibitive
    - ELK stack
    - NWperf, cview
    - Cacti - building management, facilities data

Specialized collection and vis for largescale installations?