Addressing the challenges of "systems monitoring" data flows

Mike Showerman

Jim Brandt

Ann Gentile

Themes

- 2014 – How do we collect data?
- 2015 – How do we use data?
- 2016 – How do we break down the barriers?

Is monitoring scary?

- Monitoring will destroy performance
  - My boss read that paper 10 years ago
- Can I really store all of that?
  - Its too much
  - There are so many options
- Once I have it, will I be able to generate insights?

XTREME survey says Monitoring is #1
(problem)

- Current state
  - The amount of data being collected at sites varies greatly
  - The real time processing of log and event data is fragmented
- Looking towards the future
  - Data volumes are challenging
    - High resolution power data
    - Interconnect data

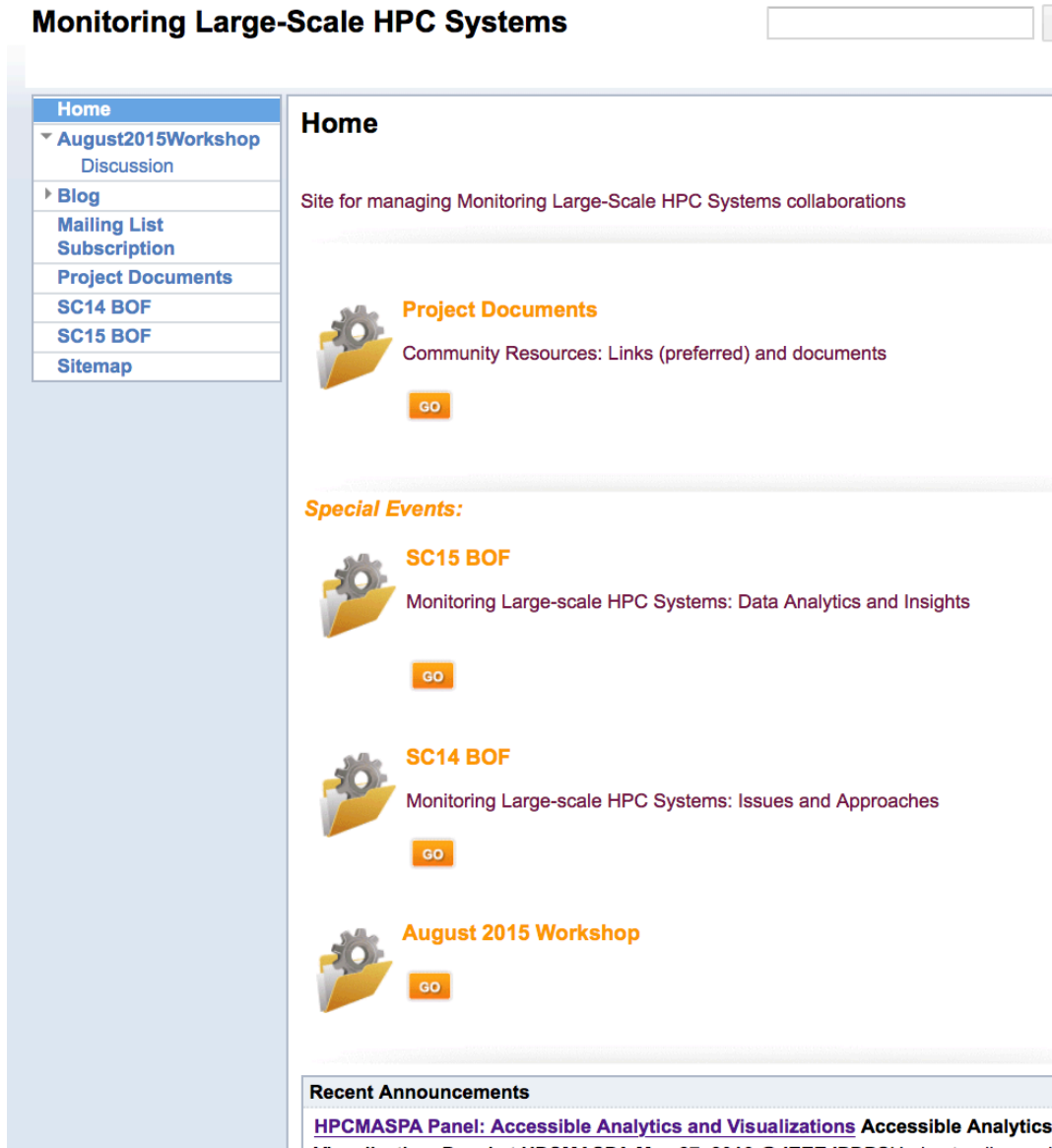
Goals of this BOF

- Understand and catalog current and future challenges
- Highlight solutions and directions
- Strengthen our interaction with Cray
  - Document our use cases for data collections

TODO's/Summary of the Monitoring BoF

Plans for making progress as a community:

- Those with functional solutions write quick start guides for the tools/technologies they are using
- Create shared list of annotated log messages
  - How can knowledge of this log line help us?
- Continue populating the Use Cases, Services, and Data of Interest document.
  - Summary of this is on the next 4 slides.
- There is a vendor/platform-agnostic monitoring community web site and mailing list at: **https://sites.google.com/site/monitoringlargescalehpcsystems**
  - Will host general community versions of these documents
  - Upcoming: Cray specific Monitoring Working being formed in collaboration with Cray. Details TBA



Diagnostics related Use Cases

- Diagnosing reason for large **apparent variation in same jobs runtimes.**
  - Other applications competing for resources
    - Messaging
    - IO
  - Failing components, full filesystem etc
- Diagnosing **job failure** and mitigation strategy
  - Resource exhaustion (e.g., memory)?
- Understanding and Diagnosing **HSN performance** characteristics (how can we tell if good or bad?) including reasons for congestion related events such as Quiesce and Throttle
- Diagnosing **file system** slowness (e.g., DVS, Lustre, GPFS)
- Diagnosing **Workload Manager** problems
- **Efficient use of logs** and alerts as diagnostic tools
  - An understanding of actual severity of an error or failure log message would enable the administrator to focus on actual problems

Understanding Related Use Cases

- How do we identify current **system bottlenecks** and trends in resource utilization to correctly design next generation systems?
- Understand **power needs** of entire HPC eco-system including cabinets and external services such as SMW/service racks and Storage
- Understand **power utilization profiles** of jobs and components
- Run-time **correctness** checking across all nodes and services
- Want full stack view from hardware to application to enable **performance** understanding in the context of system state and workload
- Data center wide “tactical” **visualization and analysis tools** (pertinent information)
  - What is running where, what resources are being consumed, what bottlenecks are being encountered?
- Understand **whole storage hierarchy** use and failure characteristics (datawarp, disk/cache, tape, etc.)

Information Access Related Use Cases

- Fundamental insights into **how things work** and a channel (person) to get that information.
  - A human interface
- **Job** information gathering using slurm to better understand job characteristics
- How are **“failures” defined** and why?
- Would like **Sonnexion** information integrated into other monitoring data streams
- **KNL** diagnostics information
- Ability to **add monitors** at will
- A high resolution **transport** mechanism
- **Published APIs** for getting at information of interest
- Need the ability to **define the level of information** from any component in the system including compute nodes

Desired Data Interaction Paradigms

- Stream data to multiple off-platform analysis targets
- Integrate data with off-platform (e.g., data center, other platforms, weather) data, both for analysis and response
- All raw data made available
- Collect from customer-defined data sources (e.g., anything on the platform or from applications)