Effective Fault Tolerance in Large Scale Computing Systems – Production Clusters

Potential Research Directions

Anwesha Das

21st February 2019

Today's Talk

- Performance Logs for HPC Application Diagnosis
  - Tools: LDMS, PerfExpert – SC'14, SC'10
  - Application: TPDS'18, ISC'18, CORRMEXT HiPC'17

- Failure Prediction Solution in non-HPC context
  - DeepView – NSDI'18 (Virtual Hard Disks)
  - Prefix – SIGMETRICS'18 (Network Switches)

- Potential Research in Large–Scale Computing Systems
  - ✔ Connection to my work (Cray–HPC, Compute Node Reliability)

Failure Prediction in Systems

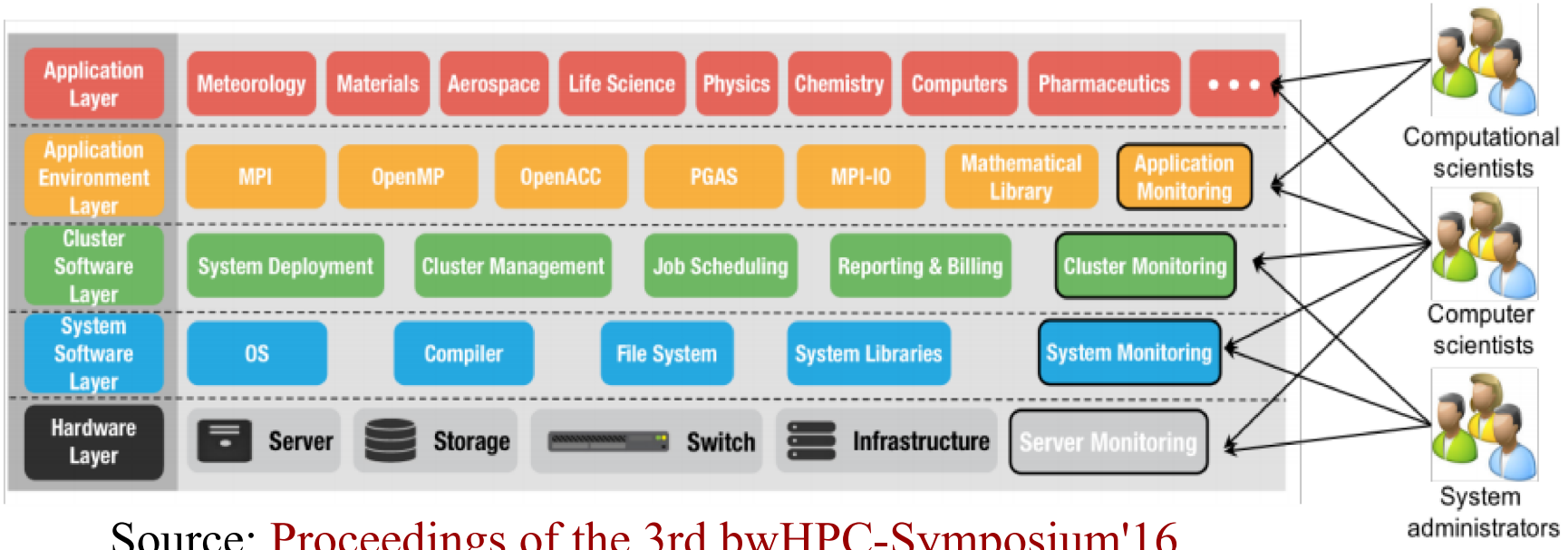
- Large–Scale Computing Systems
  - Changing Scale, Complexity, Dynamicity, Heterogeneity
  - Evolving Failures: Competing Fault Tolerance, Cascading Recovery, Performance Interference

Emergent operational behaviour unanticipated during system design

- ***Holistic*** Fault Tolerance and Recovery
  - Performance diagnosis (advanced in Clouds unlike HPC)
  - Low–level system log characterization (lots in HPC unlike Clouds)
  - Better integration of diverse components during failure analysis
  - Better co-ordination between components/layers during recovery

System–Wide Monitoring

Are the HPC clusters considered really in an integrated manner?

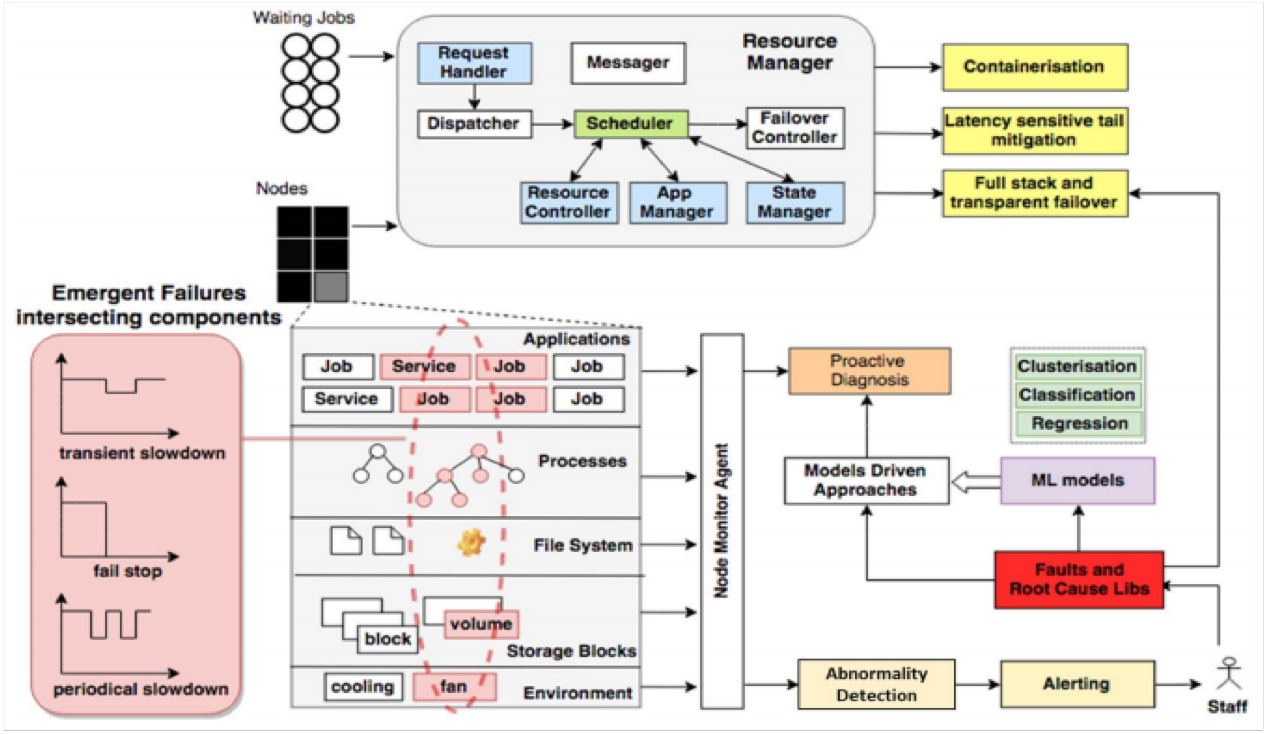


Source: Proceedings of the 3rd bwHPC-Symposium'16

- Major failure characterizations are performed focusing on the System Software Layer
- There is scope for failure analysis considering multiple layers in conjunction

System–Wide Monitoring

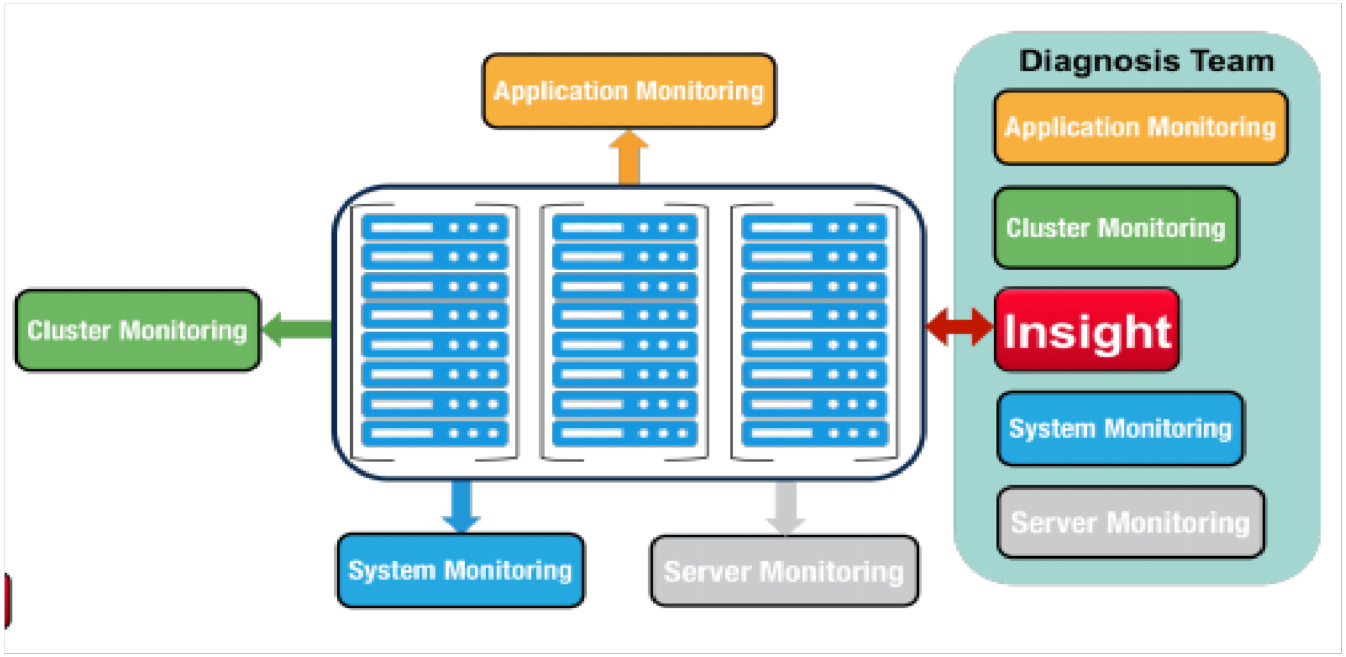
Modern data centers have focused on resource managers extensively



Emergent Failures:Rethinking CloudReliability at Scale

- Container based mechanisms, cgroup restriction
- Overall anomaly detection still not robust – scope for proactive resilience considering hardware events

System–Wide Monitoring



Source: Proceedings of the 3rd bwHPC-Symposium'16

- SEDC logs and power correlations not known yet
- Scope for further investigation: Node and job correlations with system events, resource usage data

System + Resource Usage Correlations

- Most work on Ranger Supercomputer at TACC
  - All papers by Edward Chuah et al. [HiPC'17]
  - Identify earlier resource usage anomalies prior to system faults
  - Correlations+Time Bin Extraction

- Application Resilience
  - Performance logs to raise false alarms without considering system logs

Node Failure Prediction

- **Systemic Assessment + Proactive Performance Diagnosis**
  - Less quality work currently in the literature (HiPC'17, SRDS'13)
  - Correlate performance logs with underlying system events?
  - More research w.r.t. development/engineering efforts
- **Integrated Cray Deployment**
  - Cray logger Assessment (Desh+Aarohi), Scaling out on multiple nodes
  - Addressing False Negatives/Positives, Suitable action during achievable lead time
  - ***Transition*** research prototype to production cluster
  - More development efforts, less novel research (Industrial Track or Deployment Experience)
  - ***Equally important to move from academia to industry or national labs !!***

Node Failure Prediction

- ***Improvements on existing solutions***
  - Phase 1: Static (Offline) Training
    - ✓ Performance optimization, Adaptability (many papers)
    - ✓ Unbiased model, ML Fairness (Production ML system)
  - Phase 2: Dynamic (Online) Inference
    - ✓ Ranking of nodes in terms of health for future job scheduling (FSE'18)
    - ✓ Real-time streaming logs versus static files

- ***Rethink/Improve existing node, job, system logs, resource usage correlations***
  - Known dates of soft lockup, less generic, limited correlations
  - Finer assessment of job versus node correlations for failure diagnosis

Node versus Job Correlations

- How are the job states on nodes during healthy and unhealthy times? Are there any discerning patterns?
- Are there correlations between job disruptions and node events? (not NHC logs in P0-directories but inside Torque/Slurm logs)
- Existing work LogAider (Mira RAS and Job Logs) (CCGrid'17)
- Understand coherence of unsuccessful jobs and failed nodes over time

Path Ahead

- Adaptive Fault Tolerance
  - No rigid strategy, with changing cluster composition

  (Data Center, HPC, IoT, Embedded/Real-Time, Fog)
  - Emerging failures, evolving strategy w.r.t. current and future system

  design and expected performance
  - Not threshold based, sensitive to the evolving operational context
  - Rethink system abstractions
  - Fine–grained understanding of what incidents in a system lead to

  diverse failure manifestation still requires further research