AUTOMATED SERVICE MONITORING IN THE DEPLOYMENT OF ARCHER2

Eimantas Kazakevicius, EPCC

THE UNIVERSITY of EDINBURGH

epcc

Background – ARCHER2

- ARCHER2: HPE Cray EX supercomputer
  - 5,860 compute nodes
  - Each nodes has two AMD EPYC 7742 64 core processors
  - Slingshot interconnect
  - Shasta cluster management software
  - 3x5PB L300 FS, 1x1PB nvme E1000
  - Hosted at Advanced Computing Facility, EPCC's data centre
  - Successor to ARCHER, 4,920 node Cray XC30
  - Funded and managed by UKRI

Background – ARCHER2

- EPCC provides:
  - Service Provision
    - System management and administration
    - Operation of the service desk
  - Computational Science and Engineering
    - Deployment of application software not included in the programming environment
    - Support for users with application software development/management
    - Provision of training
    - Administering funding calls
    - Outreach
  - Accommodation
    - Physical hosting and support for the system

Background – ARCHER2

- ARCHER2 experienced an extended and somewhat troubled deployment.
- Issues were faced with the development and scaling of the HPE Cray EX and Slingshot technologies.
- Given these issues the project moved to a phased transition.
- A 4 cabinet system was temporarily deployed to a separate computer room.
- This operated in parallel to ARCHER until it was possible to deploy the full 23 cabinet system.

Background – ARCHER2

- Original deployment timeline:
  - February 2020: ARCHER to be decommissioned
  - March 2020: ARCHER2 to be delivered to ACF
  - May 2020: ARCHER2 to be made available to users

Background – ARCHER2

- Final deployment timeline:
  - July 2020: ARCHER2 4 cabinet system delivered to the ACF
  - October 2020: ARCHER2 4 cabinet system made available to early access users
  - November 2020: ARCHER2 4 cabinet system made available to all users
  - January 2021: ARCHER system decommissioned and removed from the ACF
  - February 2021: ARCHER2 23 cabinet system delivered to the ACF
  - November 2021: ARCHER2 23 cabinet system made available to users

Background – Monitoring

- As discussed here automated monitoring played a key role in the deployment of ARCHER2 across the length of this extended deployment period.
- We were motivated to include this from day one by our, at that point, four years of experience working with monitoring technologies.
- Previous experience had shown benefits in reducing staff workloads, improving response time and providing insight when responding to problems.

Background - Monitoring

- EPCC manages a variety of HPC and research computing services in addition to critical support infrastructure.
- EPCC sysadmins spent a lot of time tracking the state of various systems; problem detection and diagnosis typically requires looking in multiple locations:
  - Time intensive, difficult and requires a constant wide awareness.
  - Difficult to effectively diagnose new systems where team members are typically under pressure to get things up and running in short timeframe.
- We needed a “single pane of glass” approach.

Background - Checkmk

- Originally a Nagios extension, now a Nagios derivative monitoring system.
- Many checks (both Nagios and Checkmk) available already.
  - CPU, Memory, Filesystem, Interface status etc.
- Simple to create new checks
- Very simple to add new hosts, and can alter check parameters from the central user interface
- Checkmk server first installed at EPCC in 2015. Now core to our service management for HPC services.
- Since 2015 this has allowed us to provide bespoke integrated monitoring solutions for a variety of HPC technologies.

Background - Checkmk

- In order to take advantage of data gathered by Checkmk we have also deployed a Graphite metrics server and a Grafana analytics and visualization server.
- Over time we have deployed a number of specialized checks to support our HPC services:
  - DDN controller monitoring and lustre statistic capturing
  - GPFS Cluster monitoring
  - Unplaceable/orphan job detection in PBS Pro
  - Omnipath network health status
  - Compute node status via HPCM

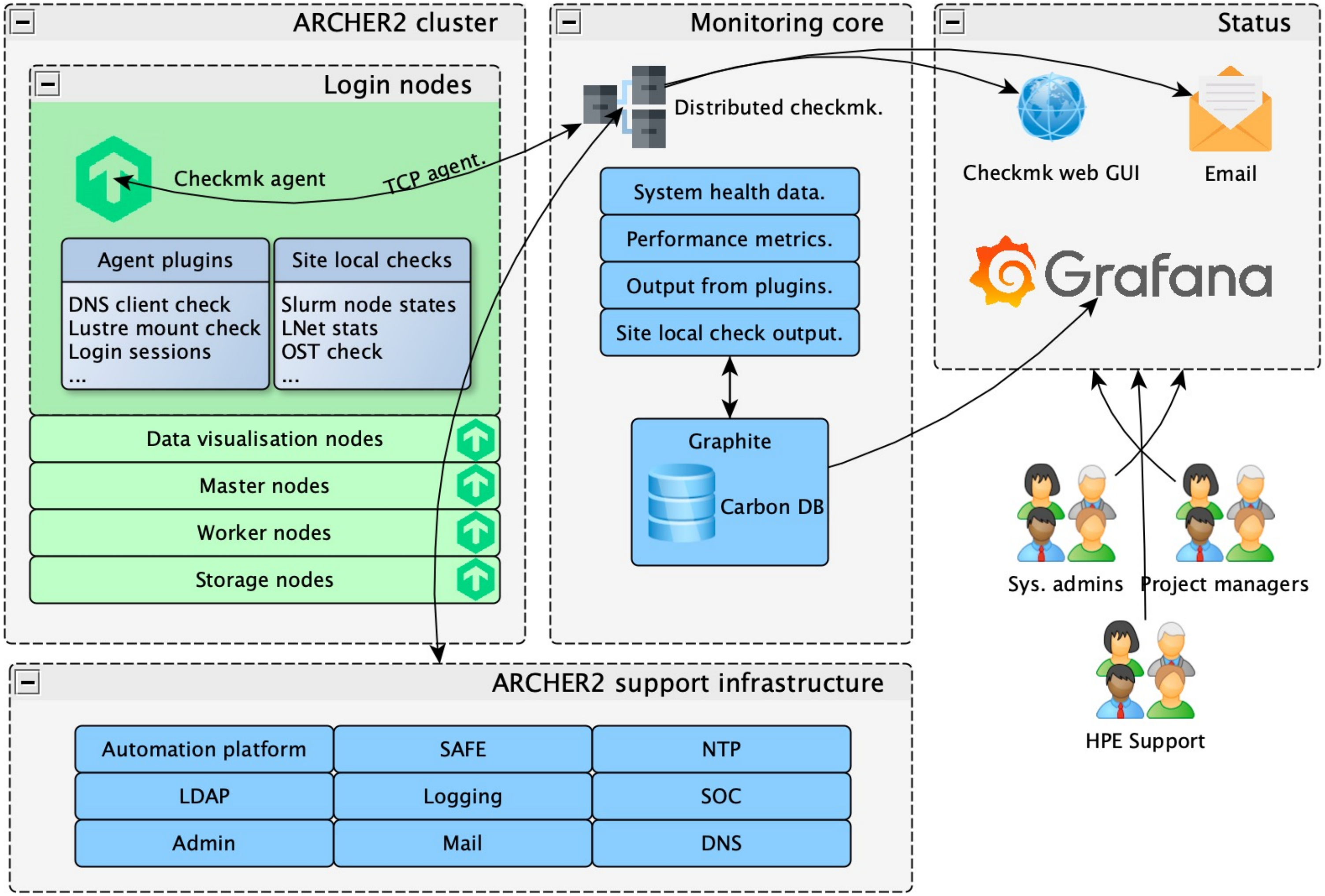
ARCHER2 Monitoring Deployment

- Separate monitoring servers are deployed for each system or group of systems.
- These are controlled from a central Checkmk instance.
- This approach has been found to improve performance and increase resiliency.
- Addition or removal of servers is simple.

ARCHER2 Monitoring Deployment

- Each monitored host has a Checkmk agent installed which communicates to the server via TCP.
- This agent collects various host health, performance metrics and posts these to the monitoring server.
- The Checkmk server passes this data to the Graphite graphing server which processes the data using ”Carbon” daemons and stores it in Graphite’s specialised database.

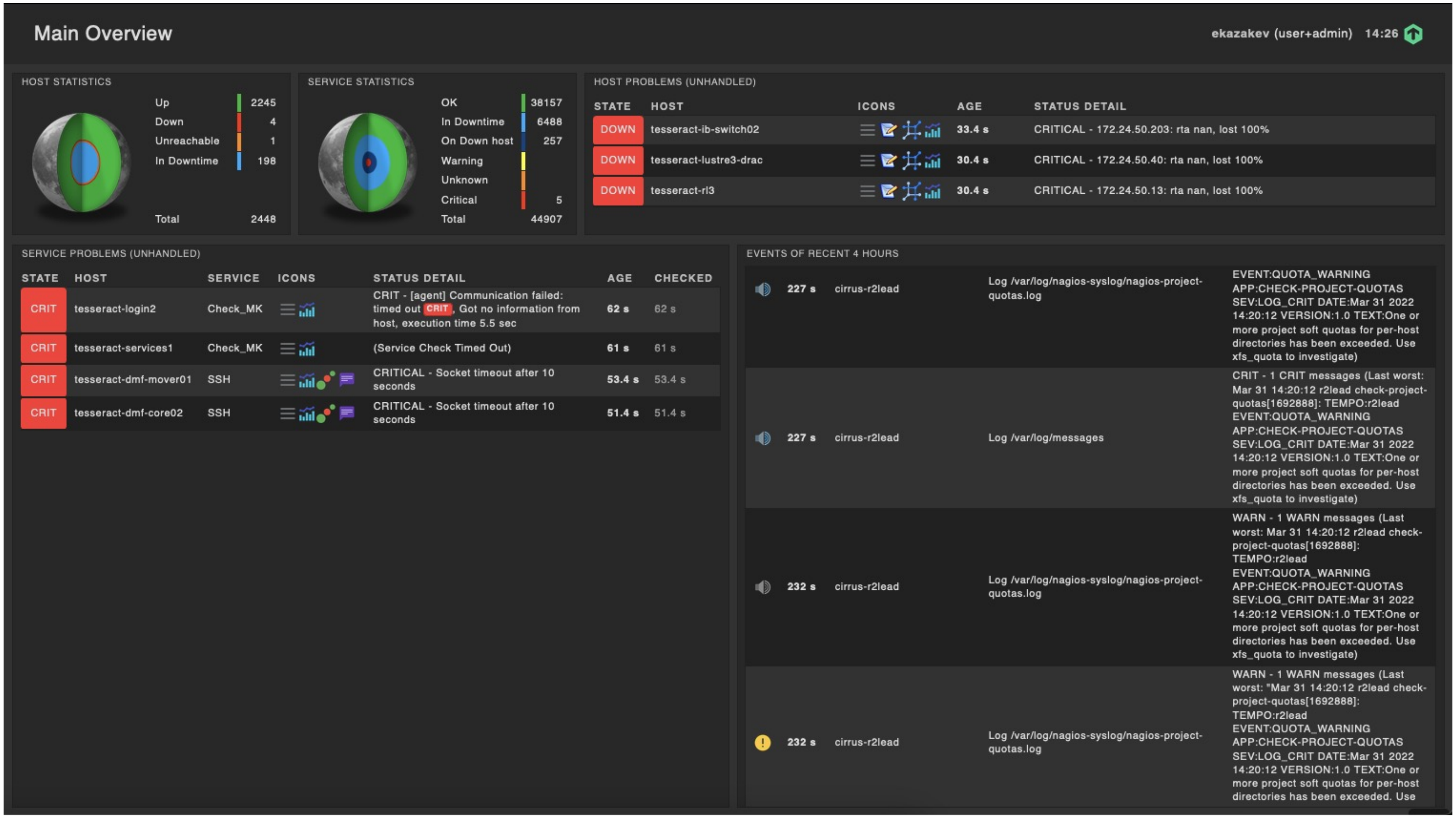
ARCHER2 Monitoring Deployment Diagram



ARCHER2 Monitoring Deployment

- Three methods to access system status information:
- All critical notifications are directly dispatched to appropriate personnel email addresses (including HPE pagers).
- Two graphical user interfaces accessible via web browser:
  - A centralised Checkmk control centre that presents overview of all hosts, services, and checks.
  - A Grafana analytics and visualisation web application that pulls various metrics from the Graphite metrics server and presents them in the form of customisable and versatile graphs.

Checkmk Front Page



Grafana ARCHER2 view

ARCHER2 Monitoring – Custom Checks

- Deployed as bash scripts placed in the appropriate directory (/usr/lib/check_mk_agent/local)
- Can be deployed using any language supported by the host.
- Only requirement is that the check output in the correct format.
- Once deployed to the appropriate directory discovery is via the Checkmk web interface.

ARCHER2 Monitoring – Custom Checks

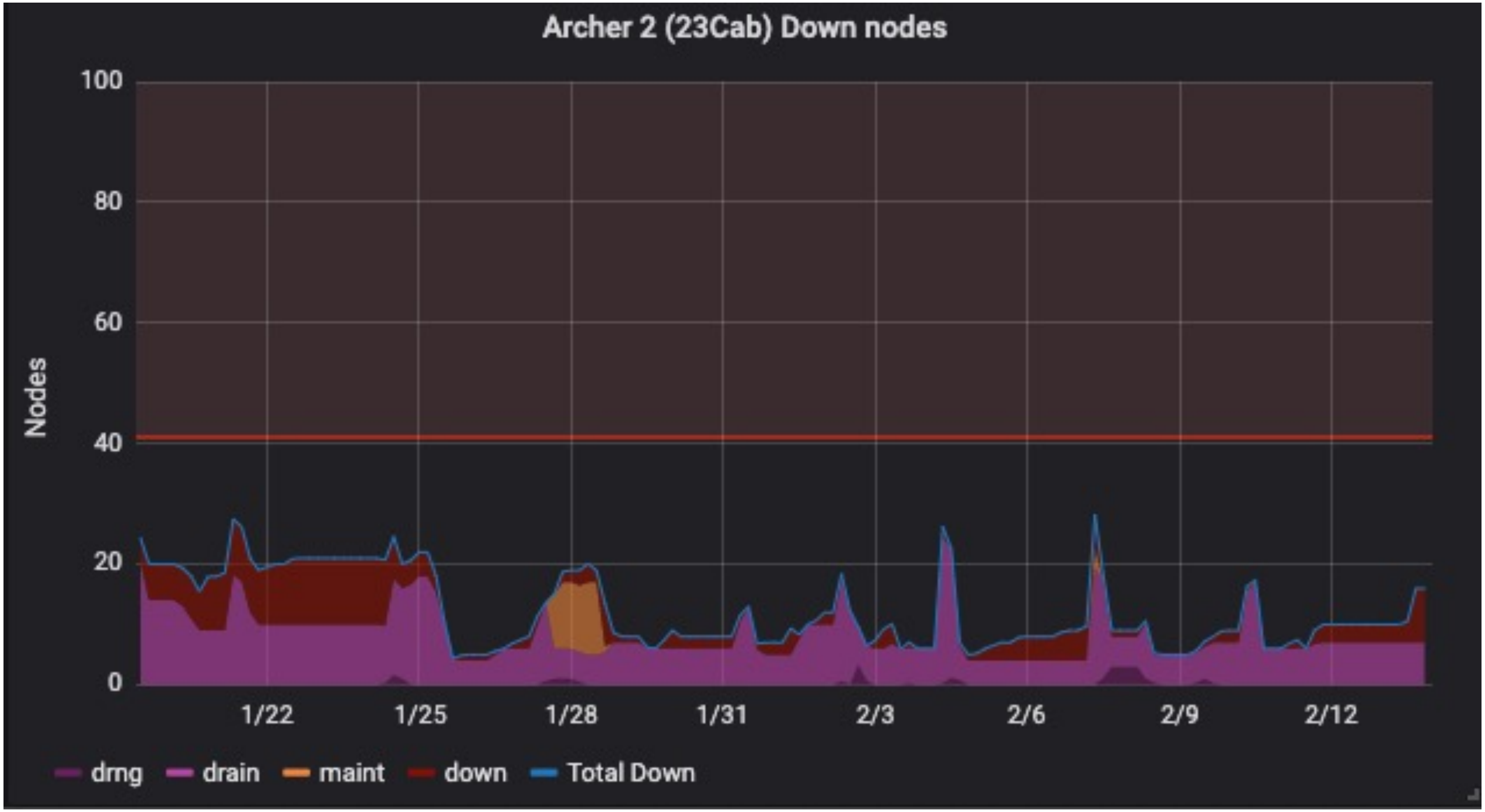
- Power monitoring
  - Runs on management node.
  - Based upon script provided by HPE.
- Process:
  - Uses pdsh to access each cabinet controller in turn.
  - On each cabinet controller gathers power data found in /var/volatile/cec/rectifiers and stores this for analysis.
  - Iterates over the data to analyse power and voltage.
  - Outputs the power draw on a per-cabinet basis.
  - Outputs the power draw on a whole system basis.
  - Outputs the voltage on a per-rectifier basis.

Power monitoring data for ARCHER2 via Grafana

ARCHER2 Monitoring – Custom Checks

- Node state monitoring
  - Runs on login nodes – clustered to support resiliency of data collection.
  - Portable – reports based on partitions listed.
- Process:
  - Runs ”sinfo” and stores the output.
  - Pulls the names of the various partitions from the sinfo output.
  - For each partition stores the number of nodes in each of the possible Slurm node states.
  - Outputs the total counts for each node type on a per-partition basis.

Node state data for ARCHER2 via Grafana



ARCHER2 Monitoring – Custom Checks

- Login availability monitoring
  - Runs on the Checkmk server itself.
  - ARCHER2 login service operates with a DNS round robin address – this check is to track whether the login service at this address is available.
  - A functional test account has single factor (key based) access available only from the Checkmk host.
- Process:
  - The script SSHes to the round robin login address with the command “exit”.
  - Based upon the exit status of this ssh command the check outputs the up/down status of the login service.

Impact – Support for Deployment

- Early deployment of monitoring was found generally useful - some specific items are worth noting:
  - A number of problems were seen with DNS – deploying a DNS resolution check allowed for rapid alerting.
  - Checkmk allowed for the rapid diagnosis of a problem with user access as being caused by network issues making a file system unavailable.
  - When experiencing problems with the Slingshot HSN the first indicator was often a drop in the number of Lustre LFS servers shown as available in the monitoring.
  - We were able to become rapidly aware of a memory leak problem. Further we were able to assess when it would become a serious problem and reboot nodes appropriately until the issue was resolved.

Impact – Initial Testing

- ARCHER2 has a noticeably larger power profile than its predecessor.
- This profile sits at the maximum of the design intent for the Computer Room - additional care was needed during initial testing.
- During the first testing (HPL@4-5k nodes) power use was monitored by observing wall level PDUs and via the Building Management System.
  - The data gathered from these sources was difficult to access and not as accurate as preferred.
- HPE identified that data was available via the cabinet controllers and made this available via a script.
  - This was integrated into our Checkmk monitoring as described previously.

Impact – Initial Testing

- This provision, verified using figures gathered from wall level PDUs and the BMS, allowed us to build confidence that the system was operating correctly and safely at scale.
- Power draw of the system was profiled while running various codes including HPL and the ARCHER2 procurement application benchmarks.
- The availability of this data also allowed us to agree remote operation of the system by HPE out-of-hours earlier in the service than would have otherwise been possible.
  - HPE's US team had access to the data and thresholds were agreed at which work would be stopped.

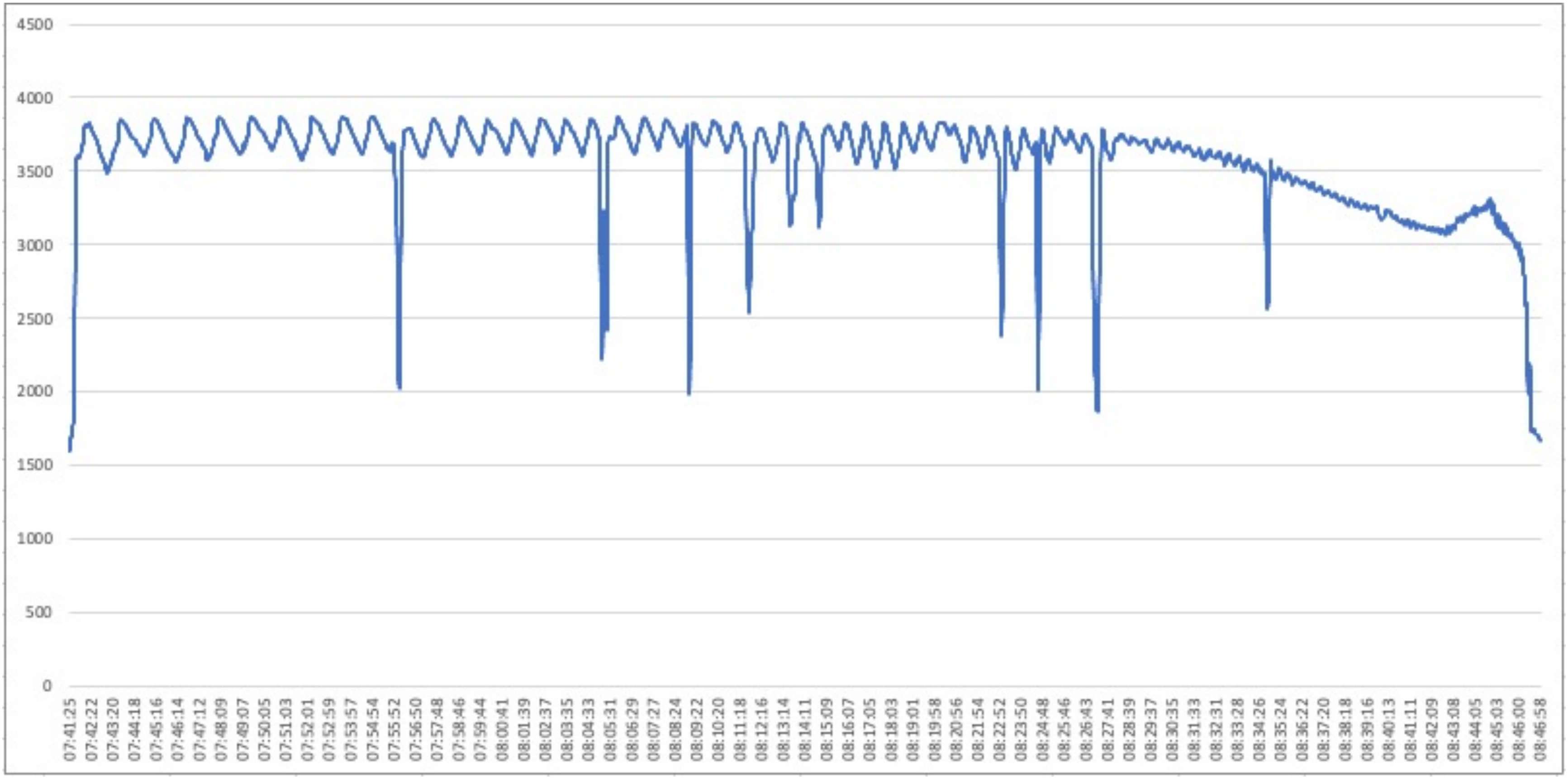
Impact – HPL Benchmarking

- Power monitoring was again useful during efforts to prepare a suitable HPL benchmark for submission to the Top 500.
- Over the course of a week a number of attempts were made to produce a suitable result – a good number of these were interrupted by node failures or HSN problems.
- Despite these interruptions we were able to complete a number of runs.

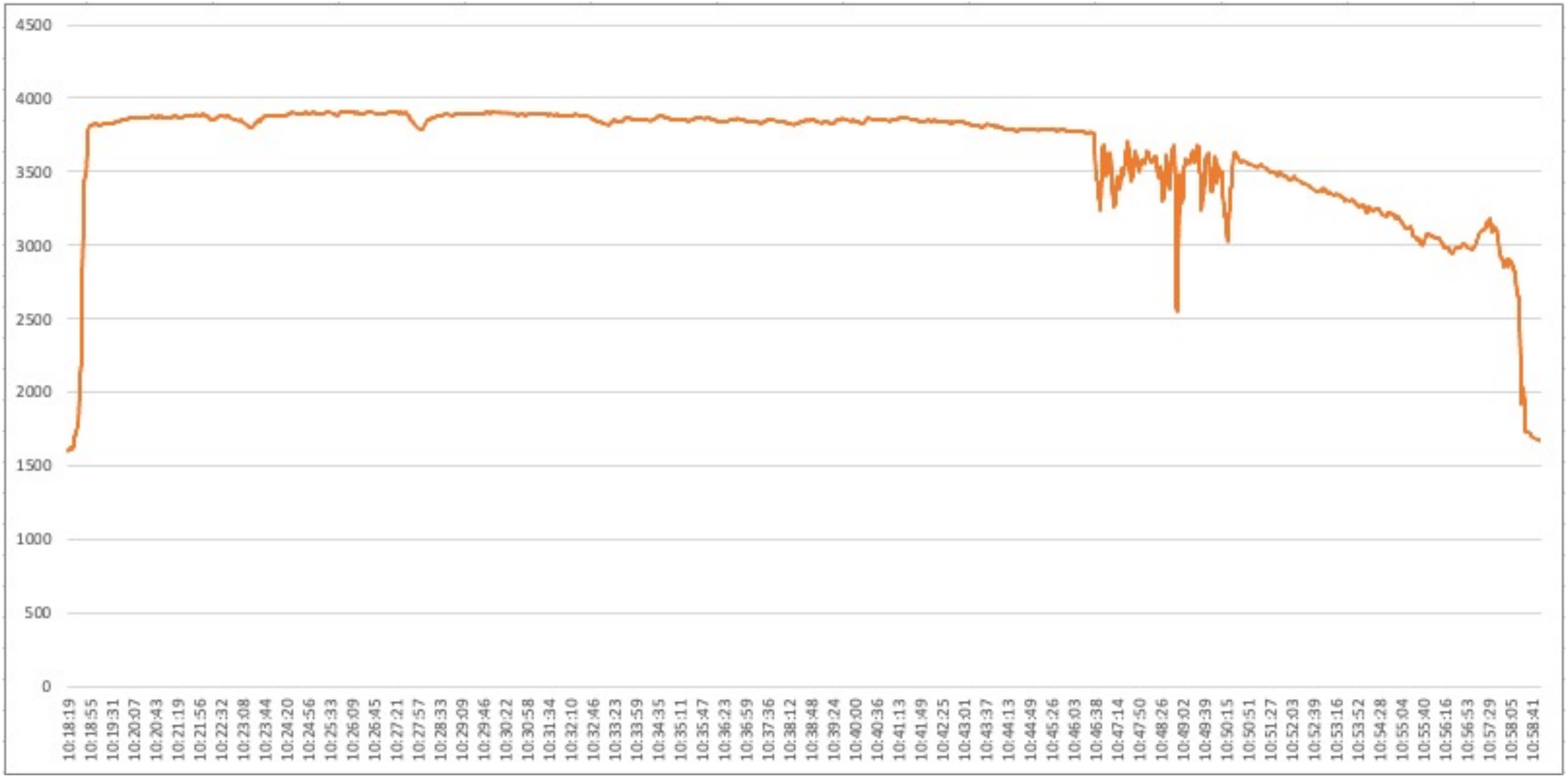
Impact – HPL Benchmarking

- It quickly became evident through power monitoring that we were seeing “power cycling” behavior.
  - Power usage would repeatedly and suddenly drop for a short period of time.
- In order to analyse this issue, single node HPL was run across the system and it was identified that certain nodes were performing persistently poorly.
- Draining these nodes removed or reduced the problem.
- This process of scanning and removing problem nodes was conducted repeatedly in order to achieve our final result of 19.5PF (placing ARCHER2 at 22 in the Top 500).

Power draw on heavily power cycling impacted HPL run (16.8PF)



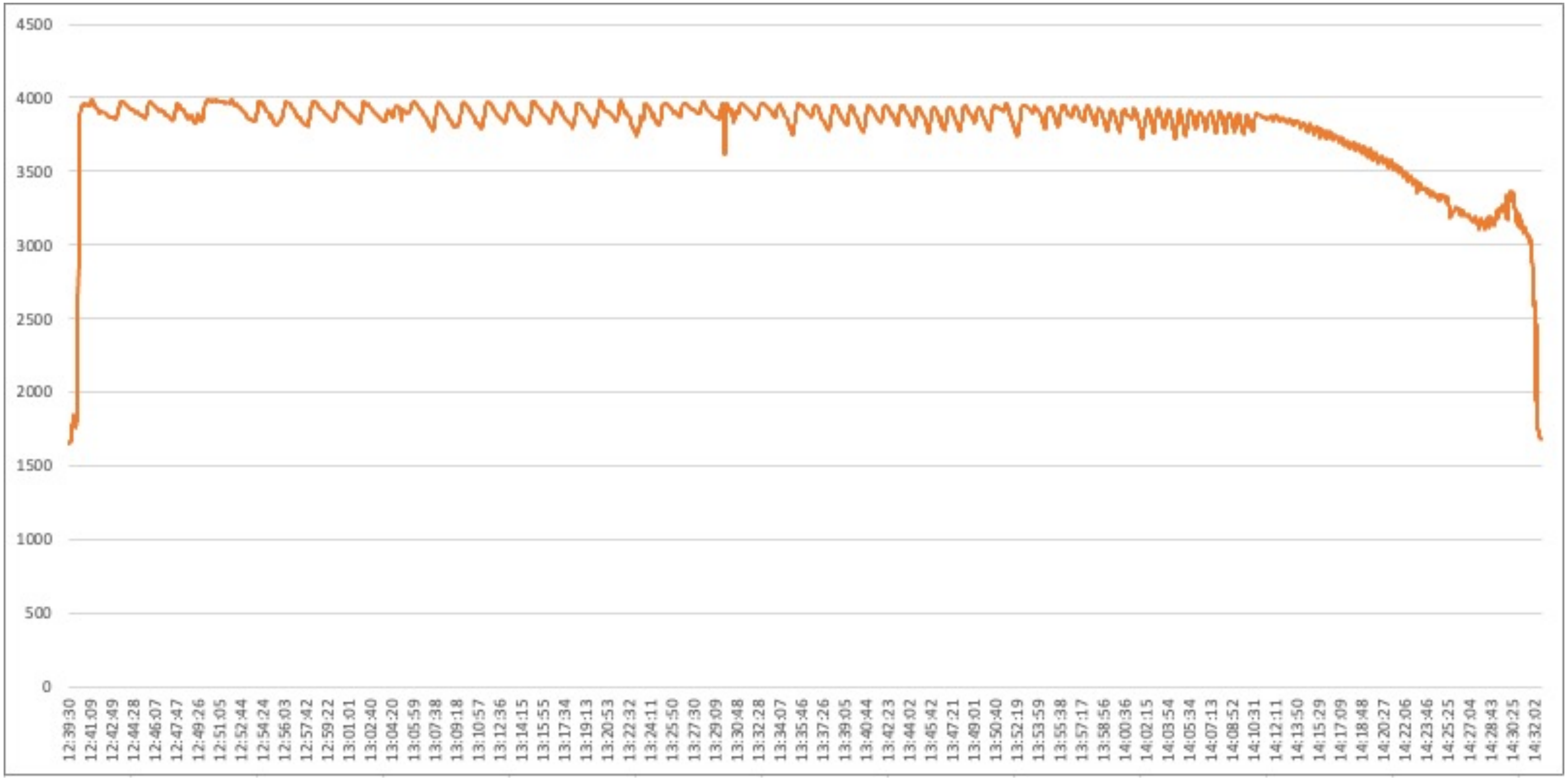
Power draw on less impacted HPL run (18PF)



THE UNIVERSITY of EDINBURGH

epcc

Power draw on submitted HPL run (19.5PF)



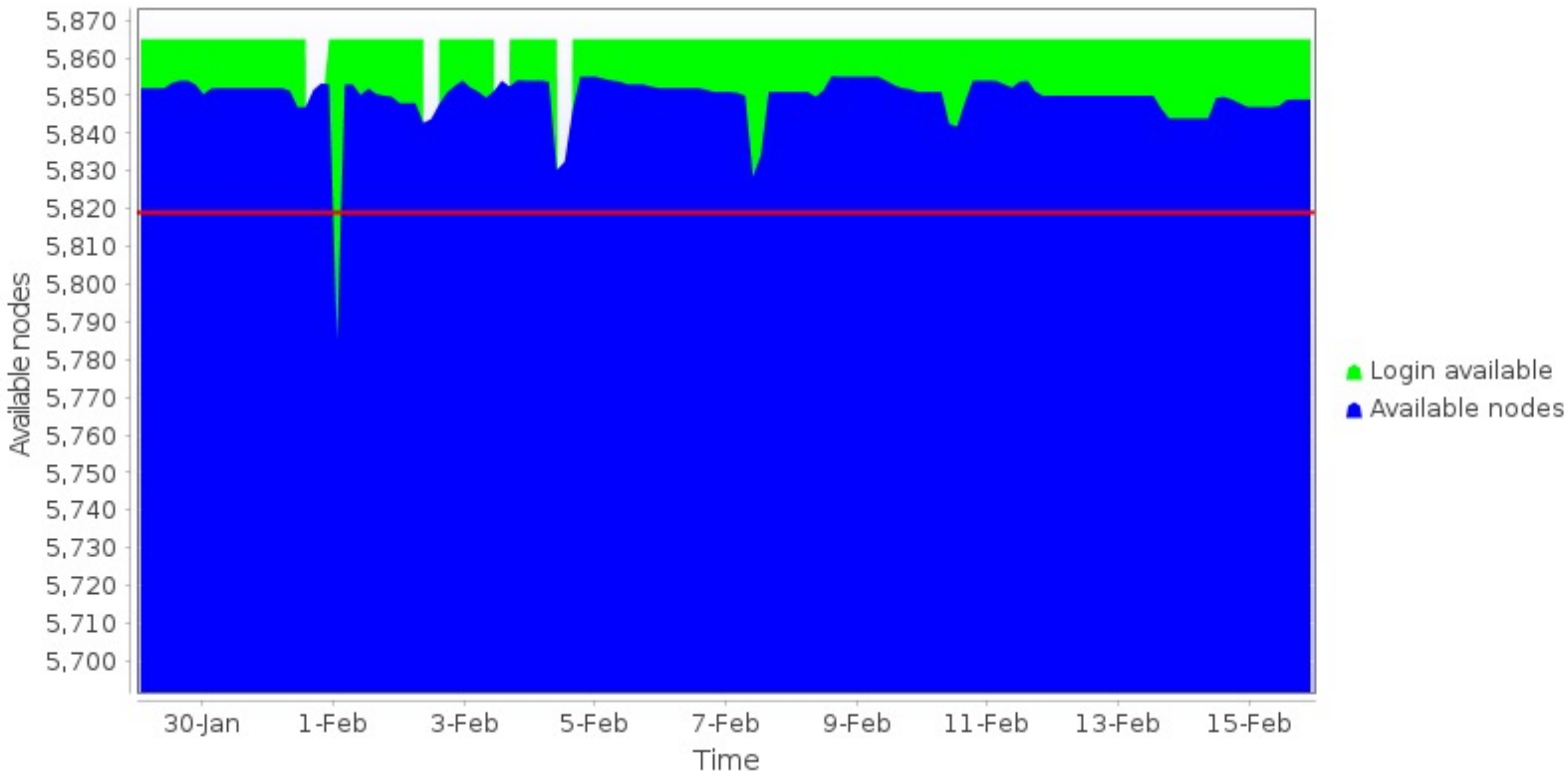
Impact – Contractual Monitoring

- In order to support UKRI (the funders) in monitoring the service during the acceptance trial a requirement emerged to present a single view of all service attributes relevant to contractual monitoring.
- The key items here were node availability, login availability and job failures.
- Data from Graphite was exposed to EPCC's service management web application, SAFE via web API over HTTP.
  - SAFE also receives all Slurm accounting and failure data.

Impact – Contractual Monitoring

- Using this data any authorised stakeholder can generate a report in SAFE covering contractual monitoring for any given period.
- SAFE provides fine-grained access control so only appropriate stakeholders can access this data.
- In addition to stakeholders in EPCC, HPE and UKRI graphing of the status/utilisation of nodes is made available on the ARCHER2 status webpage.

Contractual monitoring graph from SAFE



Future work

- Potential improvements to ARCHER2 monitoring include:
  - Log analsysis
  - Slingshot error feeds
  - Per-job lustre stats
  - Data driven intrusion detection.
- We are also interested in making the data we collect more generally available to our user community.
- We would be pleased to coordinate with other sites who use or are interested in using Checkmk for HPC service monitoring and are happy to share our experience.

Conclusions

- Live monitoring and graphing makes an extremely valuable contribution to service management.
- Value often presents itself in unexpected ways.
- The ability to rapidly and flexibly deploy new checks in response to emerging events and requirements is also of particular value.
- An imperfect check implemented rapidly is often superior to an ideal check later.
  - You lose 100% of data you don't collect (apologies to Mr Gretzky)

Conclusions

- Automating the contractual monitoring of a service can be extremely valuable.
  - Helps us to assure service partners, funders and users that system is working correctly.
  - This has been particularly important given the delayed start to ARCHER2.
- ARCHER2 has now been in full service for almost six months with in excess of 2,500 active users and utilisation on the order of 90%.
- We consider automated monitoring to have been key in making this possible.