Action Items

Communication/Site Monitoring

Frederick Luehring
April 15th
WBS 2.3 Facility Coordination

Introduction

- After a recent low CPU utilization event at SWT2_CPB, Rob asked me to suggest actions to improve the situation.
  - I will not discuss details of the event at SWT2_CPB so I can focus on doing better.
- I will also take a look at using current US ATLAS team members based at CERN to help with the US-CERN communication. (Also Rob's suggestion.).
  - This is just some trial ideas of my own and not an official proposal to make changes.

Thanks to Ofer for his help in editing/improving the talk!

Any remaining errors/problems are my responsibility.

Actions

Actions - Communication

- AGIS should be modified to send automatic notifications to site administrators when AGIS parameters are changed.
  - Changes were made without all affected parties being informed
  - Changes were made without the site doing tests that they were OK
  - It is a worry that this will overload the site administrators with too many emails.
- US-based site administrators should have regular, direct communication with EU-based ADC/ADP team members.
- Rob comments:
  - We should collect action items from weekly ADC TCB and Ops meetings.
  - How do we get the action items from ADC Dailies?

Actions - Monitoring

- We should define a single, concise set of monitoring plots/checks to ensure that all sites are functioning properly.
  - “We” means all parties: the sites, the ADC monitoring team, ADC production team, WBS 2.3.5 Operations team, etc.
  - The monitoring should be as lightweight because the site administrators are busy.
- All sites should systematically monitor/check the site each working day.
  - Don’t let things fall through the cracks.
- Sites should ensure that the necessary monitoring will remain fully available after making major changes.
  - Don’t change a major site element (software, middleware, queuing) without having the ability to fully monitor the site afterwards.

Action: Operation

- All changes affecting a site by external groups should be scheduled in advance with the site's administrators and announced widely.
- Document all of the systems for monitoring site and production system issues, tracking those issues, and communicating those issues.
  - There are an enormous number of tools - so many that I fear none of them are used.
    - E.g. there are at least 4 ticketing systems (BNL, GGUS, ADC JIRA, and OSG) .
    - Monitoring is the worst case here: no matter how many monitoring and accounting systems I find, there always seem to be more.
  - We need a standard operating procedure so site administrators use the tools effectively.
- The role of the US ATLAS ops team needs to evolve and be clearly defined.
  - The ops team should be proactively knocking on doors when there are issues.

Action - Technical (noted for the record)

- The maximum time before a Rucio transfer fails should be reduced.
  - Timeouts longer than 10,000 s occur on transfers that would have taken a few minutes if they succeeded.
  - The timeout varies widely from file to file with the average being ~4000 s to ~5000 s
  - It would be best to have a short timeout or to set it based on network speed/file size.
    - Ofer pointed out to me that we need to be careful when the file is being staged from tape with the associated long delay waiting for the tape to mount.

Working with US ATLAS people at CERN

US ATLAS People at CERM

- Clearly US ATLAS people based at CERN are in the best position to facilitate communication with the ADC teams.
  - They have the tightest connection to CERN-based groups with many personal friendships.
  - They are also in the right time zone.
- So how can we improve the situation using the CERN-based team members (given what we have learned at CPB)?
  - In that event changes were made without the necessary communication including a situation where it appears that the US and EU people were making conflicting changes to AGIS. Changes were made, reversed, made again, reversed again…
  - Apparently nothing was done to check that the site was not adversely affected. Several changes in AGIS that broke the site were made with no testing that things were working.

What We Might Do....

- Since US ATLAS has people at CERN, they should work more closely with US ops team to ensure that needed information flows both ways.
  - The CERN-based members can work with the various teams making changes that affect the production system. I believe they already osmotically have this knowledge...
  - The US-based members have good connection to the US sites and their system administrators. Again I believe they already osmotically have this knowledge...
- Particularly when a big change occurs (e.g. to Rucio mover), they should work together to announce, plan, and track the changes.
  - The CERN team learns what is changing and conveys that to the sites.
  - As sites make changes and issues occur, the issues are fed back to the developers.
- Of course the US and CERN-based people also should be working together everyday to ensure good communications in both directions.

Final Word

- It seems to me that the key here is focused teamwork.
  - The US & CERN teams should jointly plan for upgrades and solve day to day issues.
  - The US-based team should be monitoring the sites everyday and immediately questioning the sites about about things that don't seem correct. They should also be telling the sites what changes are in the ADC pipeline.
  - With the information from US-based team, the CERN-based team should get the issues resolved and, importantly, give feedback to the developers.
  - If we fall behind or have many failures, together they get things back on track.
  - We have the people doing this at some level but I would make it more formal.
- Given how busy the site administrators are, this will provide extra eye for spotting problems and getting quick answers.