Sandia National Laboratories

A Preliminary Report on Red Storm RAS Performance

(RAS = Reliability, Availability, and Serviceability)

Jon Stearley, Robert Ballance

{jrstear,raballa}@sandia.gov

Sandia National Laboratories

May 8, 2006

Cray Users Group (CUG-06)

Are these numbers Comparable?

Sandia National Laboratories

Systems	CPUs	Reliability & Availability
ASCI Q	8,192	**MTBI: 6.5 hrs.** 114 unplanned outages/month. ◆ HW outage sources: storage, CPU, memory.
ASCI White	8,192	**MTBF: 5 hrs. (2001) and 40 hrs. (2003).** ◆ HW outage sources: storage, CPU, 3rd-party HW.
NERSC Seaborg	6,656	**MTBI: 14 days. MTTR: 3.3 hrs.** ◆ SW is the main outage source. **Availability: 98.74%.**
PSC Lemieux	3,016	**MTBI: 9.7 hrs.** **Availability: 98.33%.**
Google	~15,000	**20 reboots/day; 2-3% machines replaced/year.** ◆ HW outage sources: storage, memory. **Availability: ~100%.**

MTBI: mean time between interrupts; MTBF: mean time between failures; MTTR: mean time to restore

Source: Daniel A. Reed, UNC (via Chung-Hsing Hsu, LANL)

Sandia National Laboratories

What is “RAS”???

“MTBI” **“MTBF”** **“MTTR”**

“UP” **“DOWN”**

Problem:

We DO NOT agree on terms!

This prevents accurate comparisons,

obscures meaningful discussion,

and delays significant improvements.

Solution:

Let’s agree on definitions,

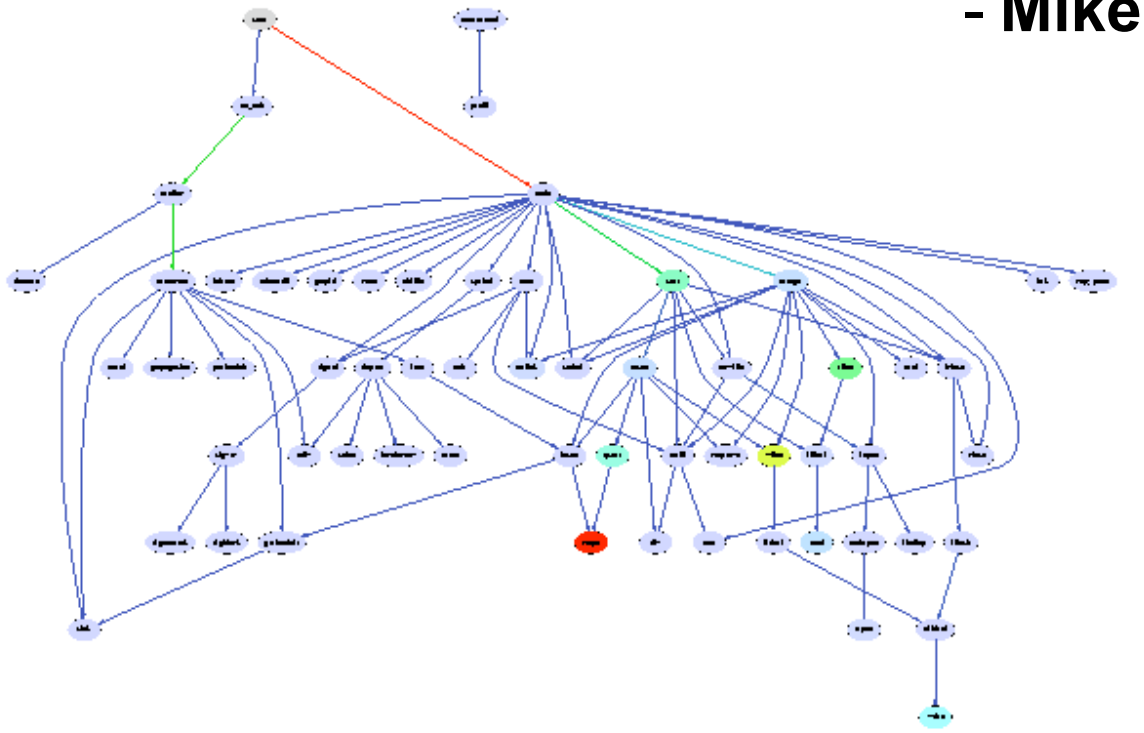
And develop a reference implementation.

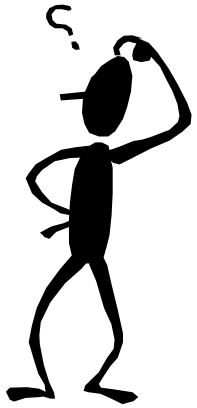
(Efforts are underway to establish a SNL/LLNL/LANL-endorsed “specification for defining and measuring high performance computing reliability, availability, and serviceability”.)

Up or Down?

"A computer is in one of two situations. It is either known to be bad or it is in an unknown state."

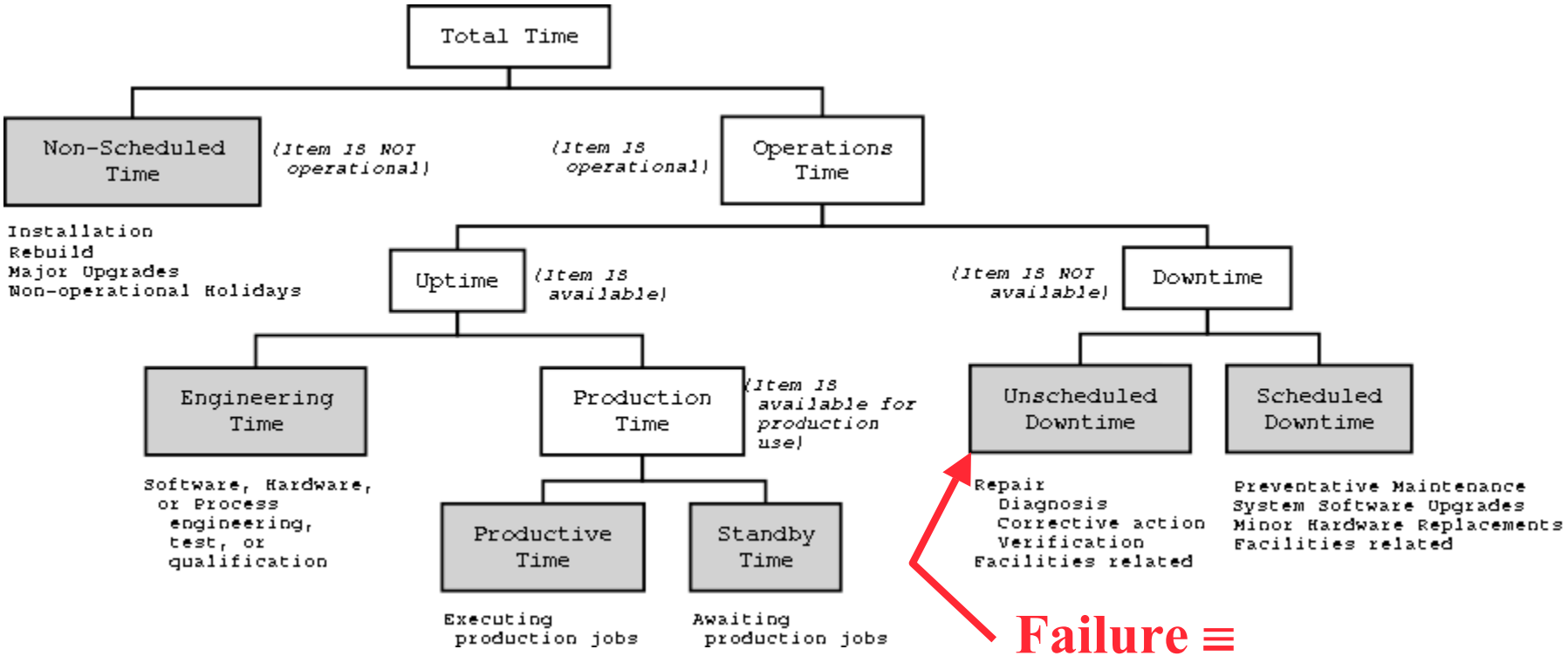
- Mike Levine (PSC)





Is the "system" "up" (yet)?

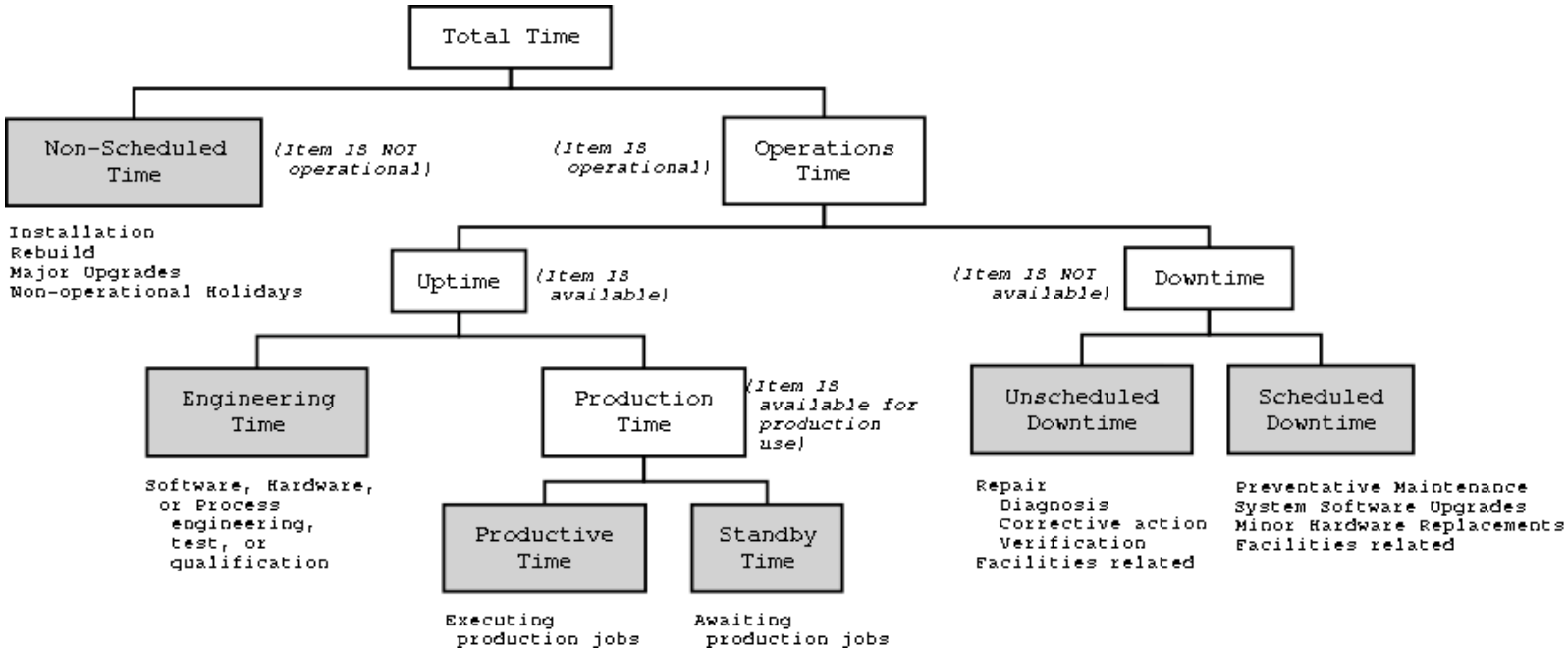
State Model
(adapted from SEMI-E10)



Items are <u>always</u> in one of the six basic states (shaded).

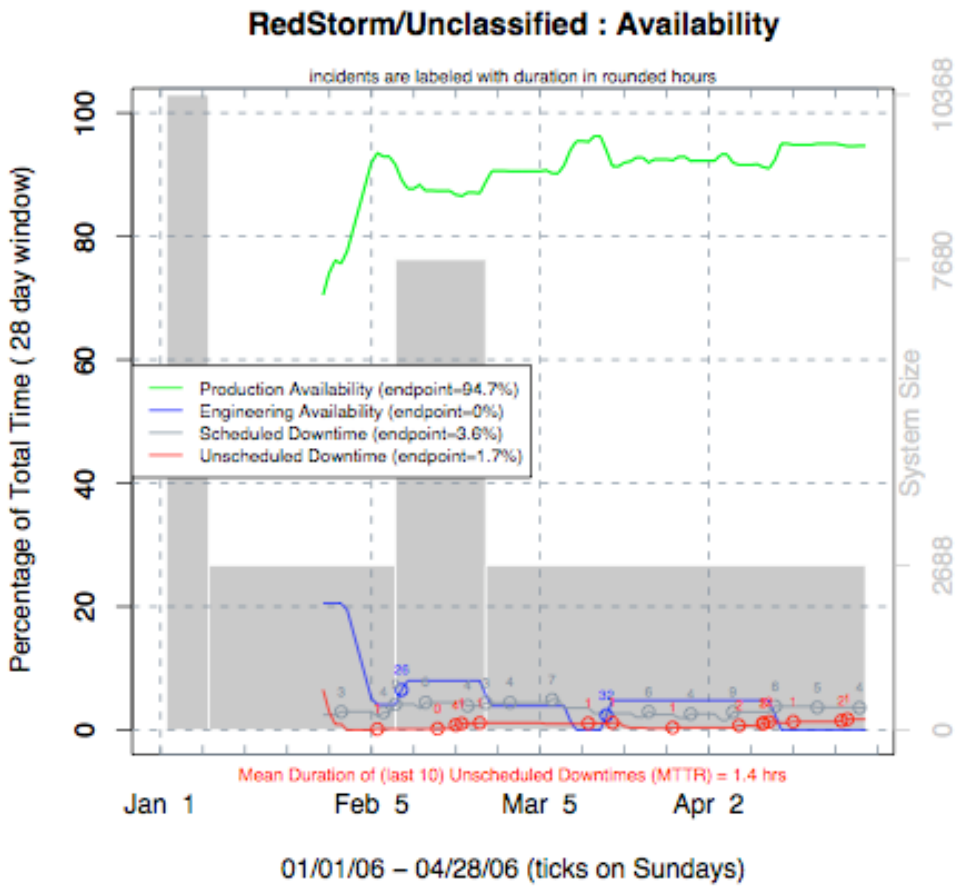
Availability

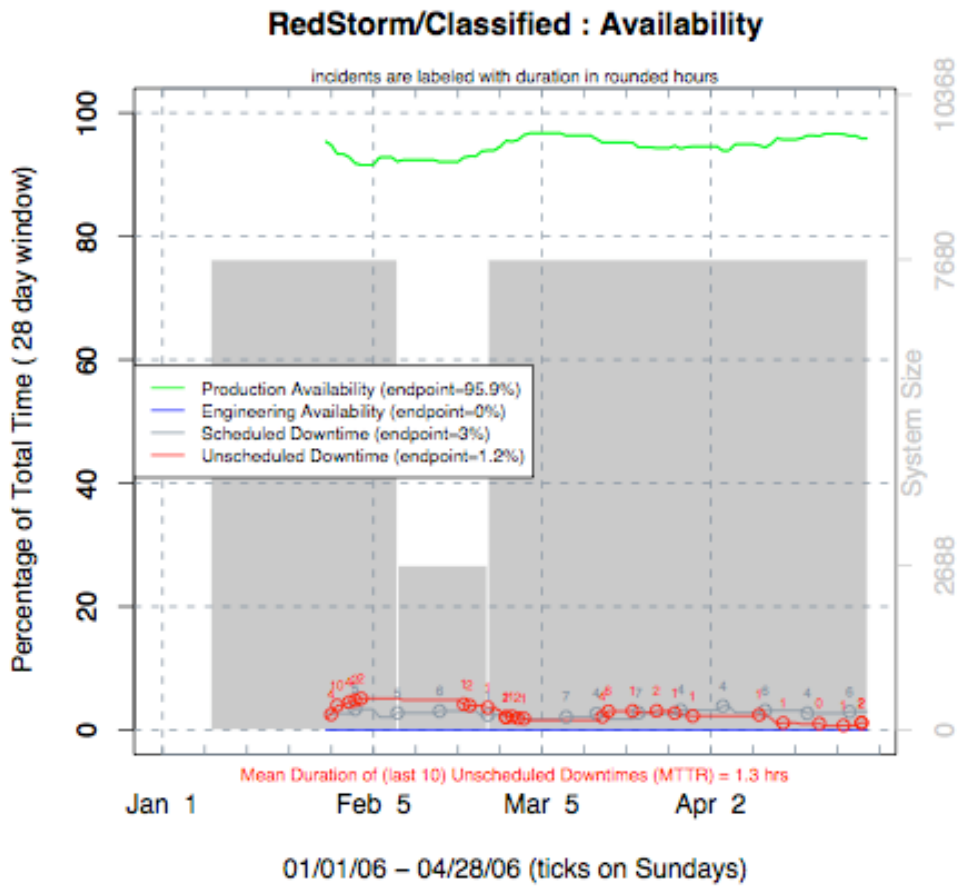
The fraction of a time period that an item is in a condition to perform its intended function ("available"). [IEEE]



$$\text{Production Availability}\,(\%) = \frac{\text{Production Time}}{\text{Total Time}} * 100$$ [SEMI-E10]

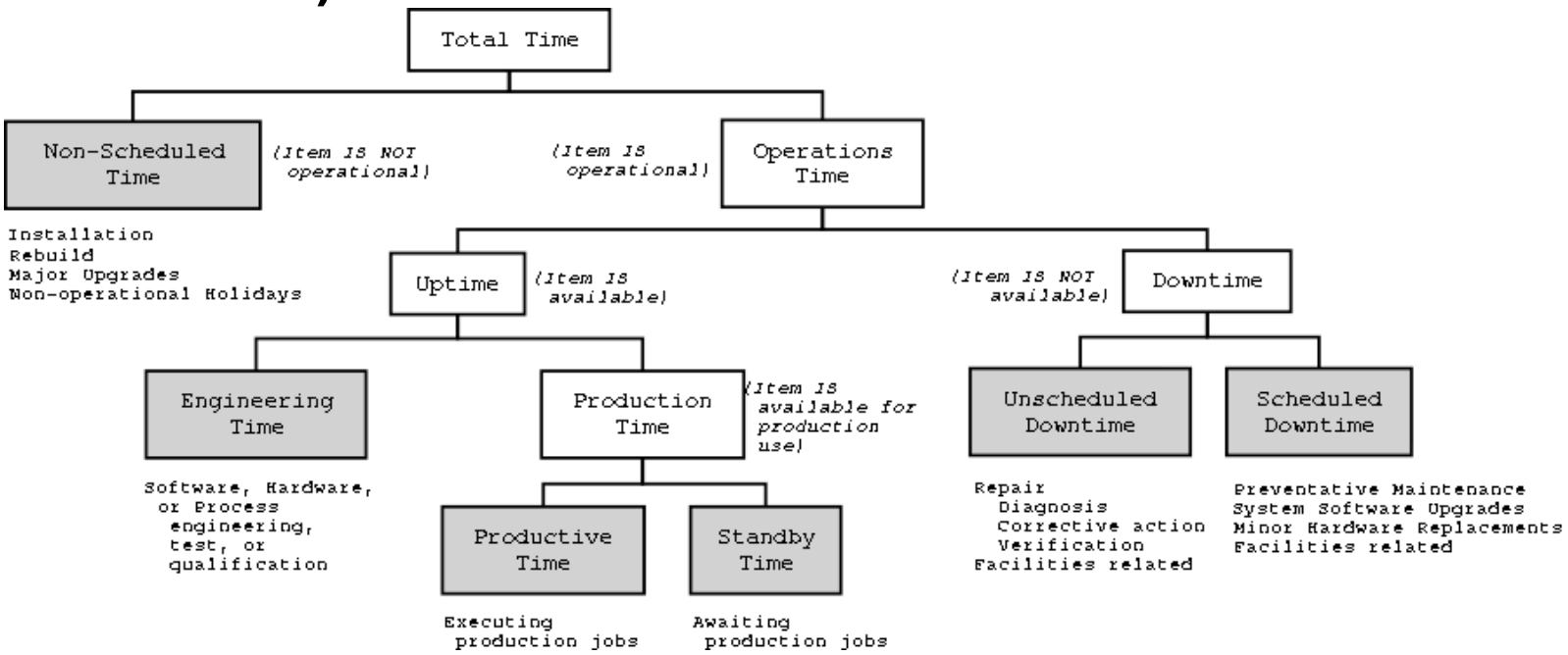
System Availability





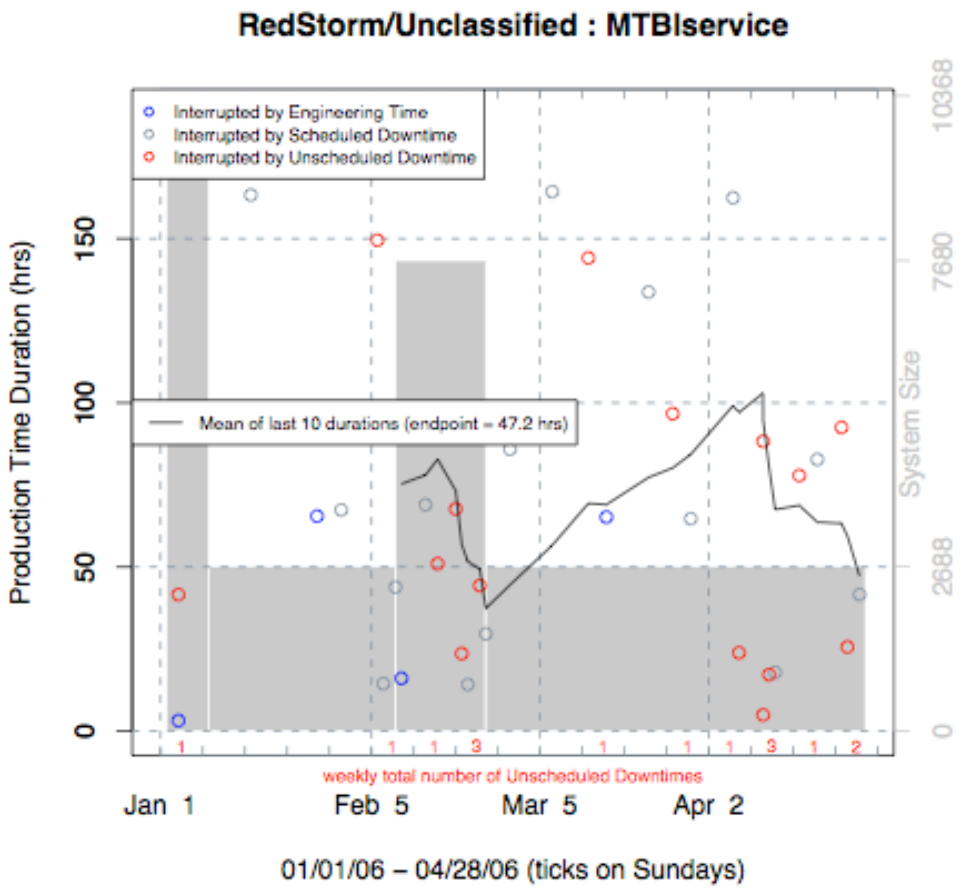
MTBIservice

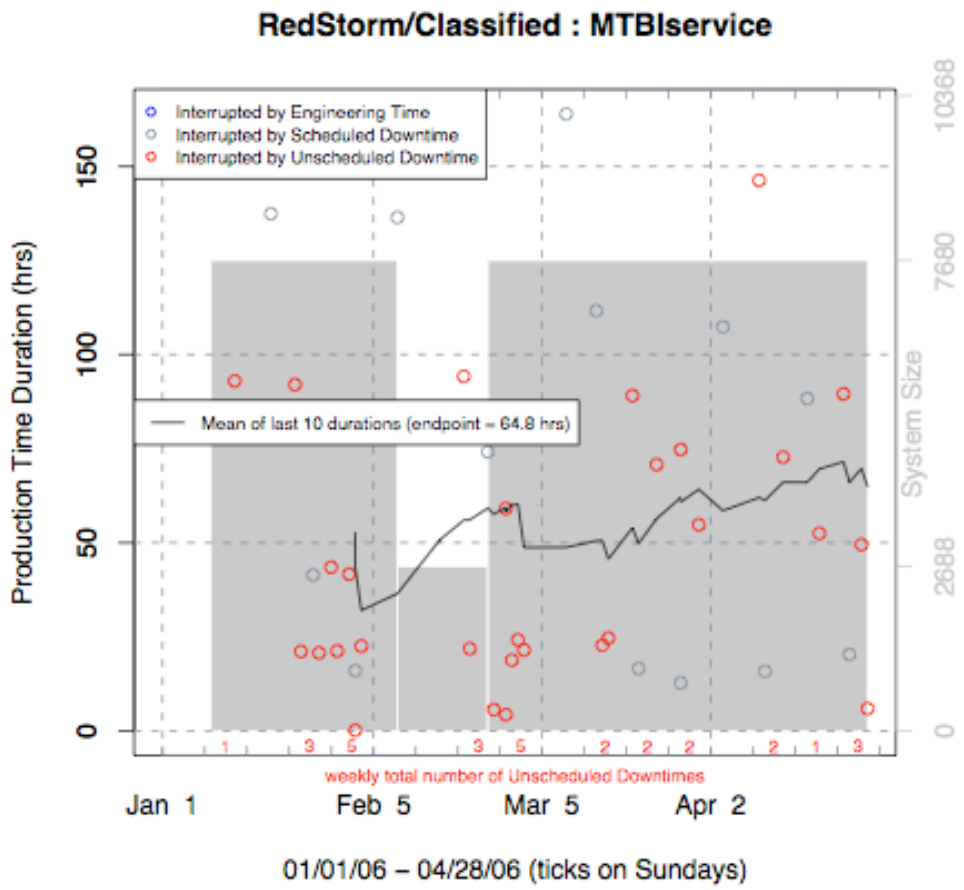
A measure of reliability describing how long the system stays in a production state (regardless of why an interruption in production service has occurred).



$$\text{MTBIservice (hours)} = \sum \frac{\text{Production Time}}{\text{Service Interrupts}}$$

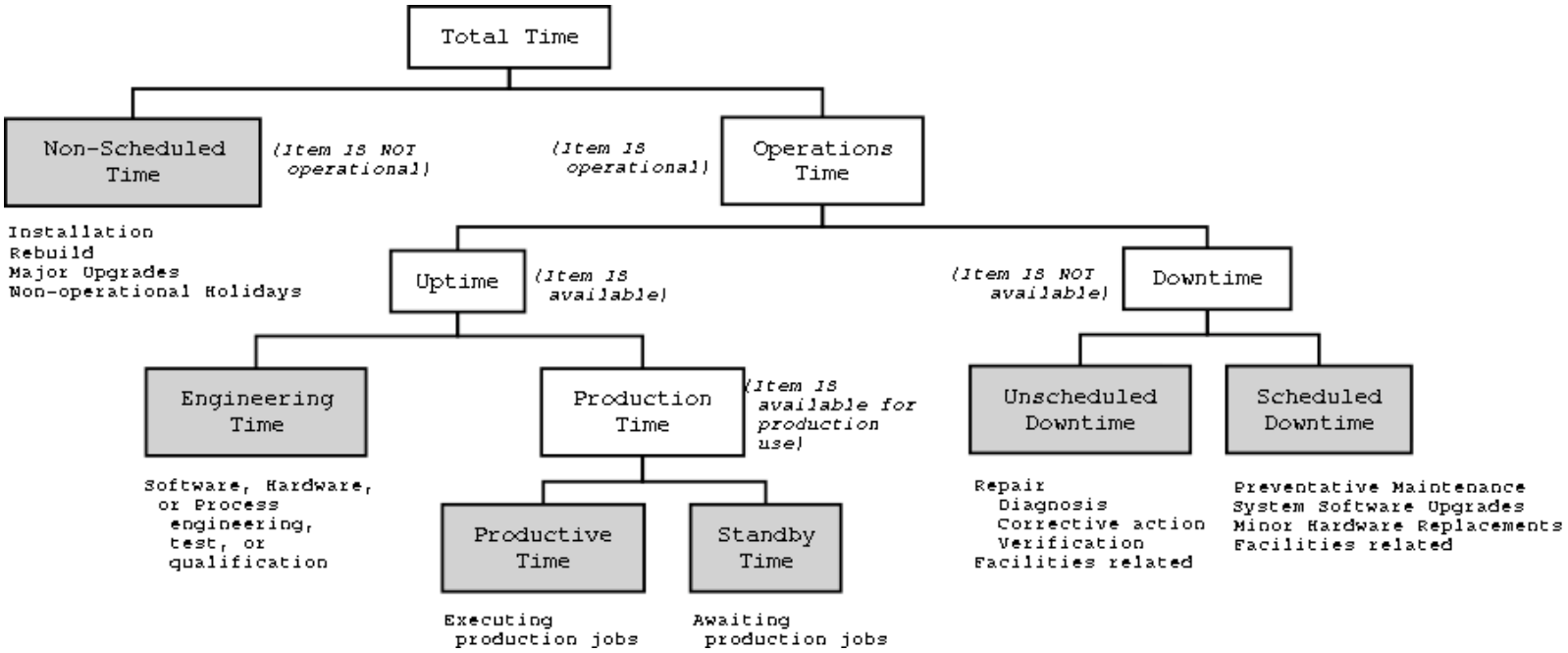
MTBIservice



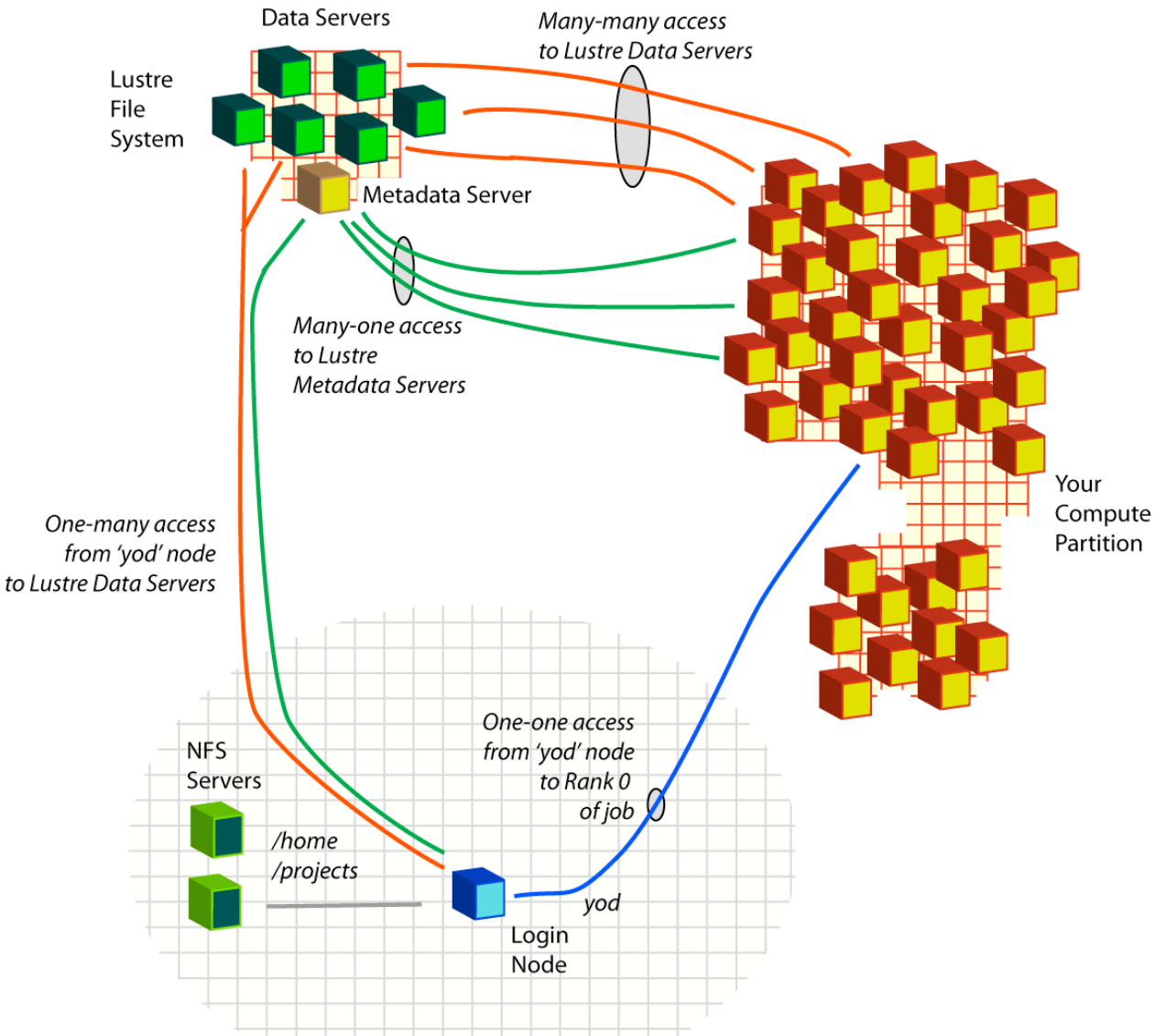


MTBIjob

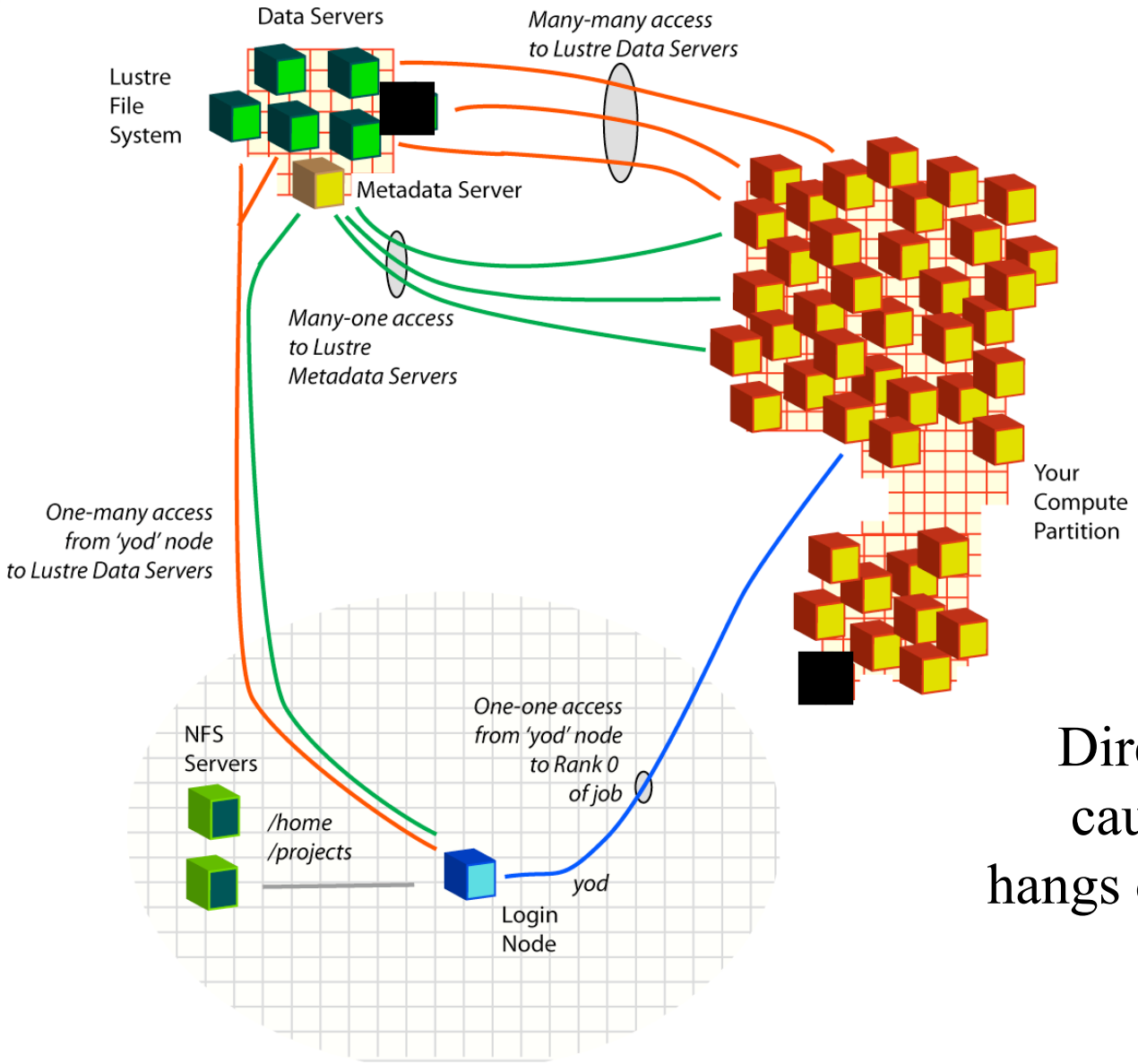
A measure of reliability describing how long the system is in a production state before any job is interrupted (by the system).



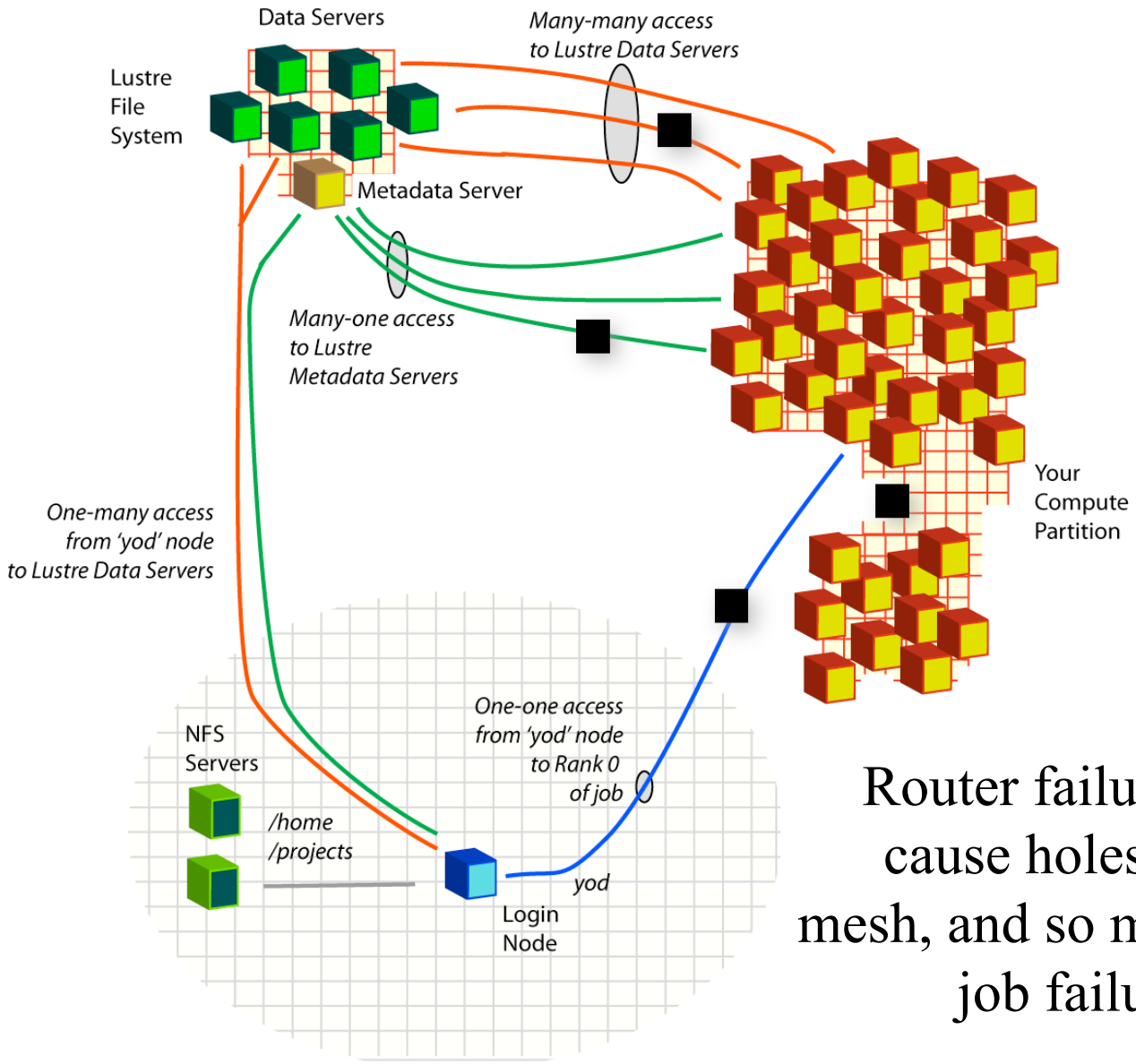
$$\text{MTBIjob (hours)} = \sum \frac{\text{Production Time}}{\text{Job Interrupts}}$$



How do you detect a job failure in an XT3?

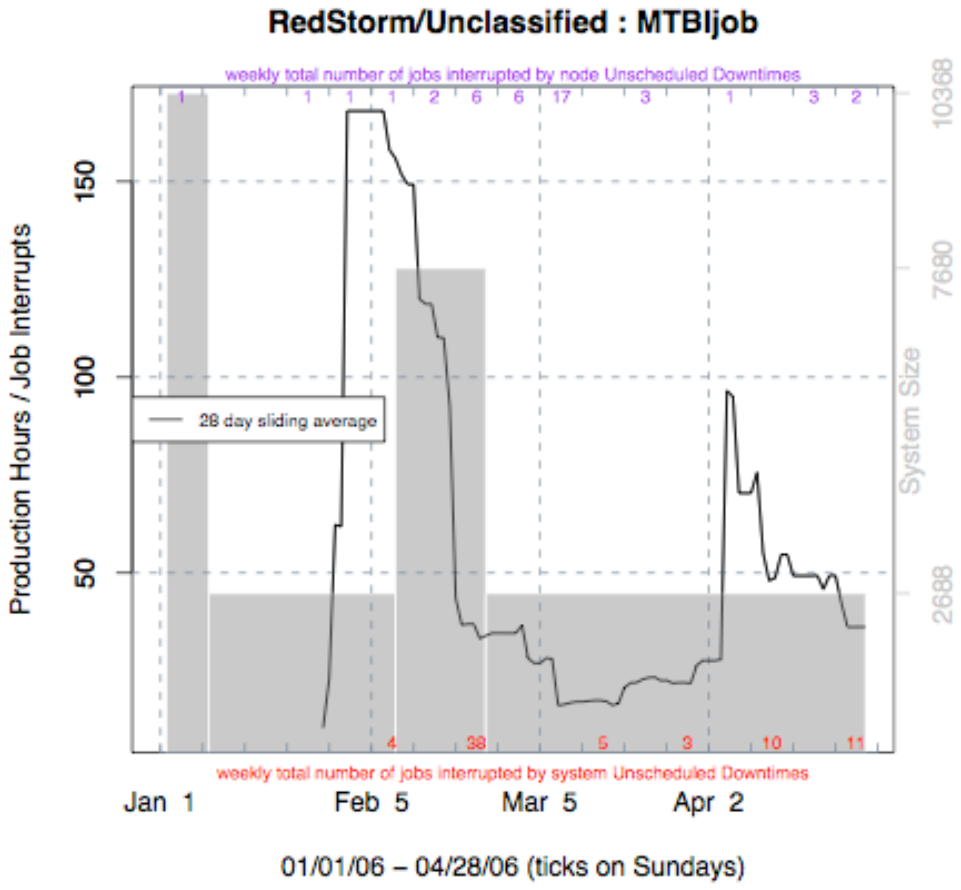


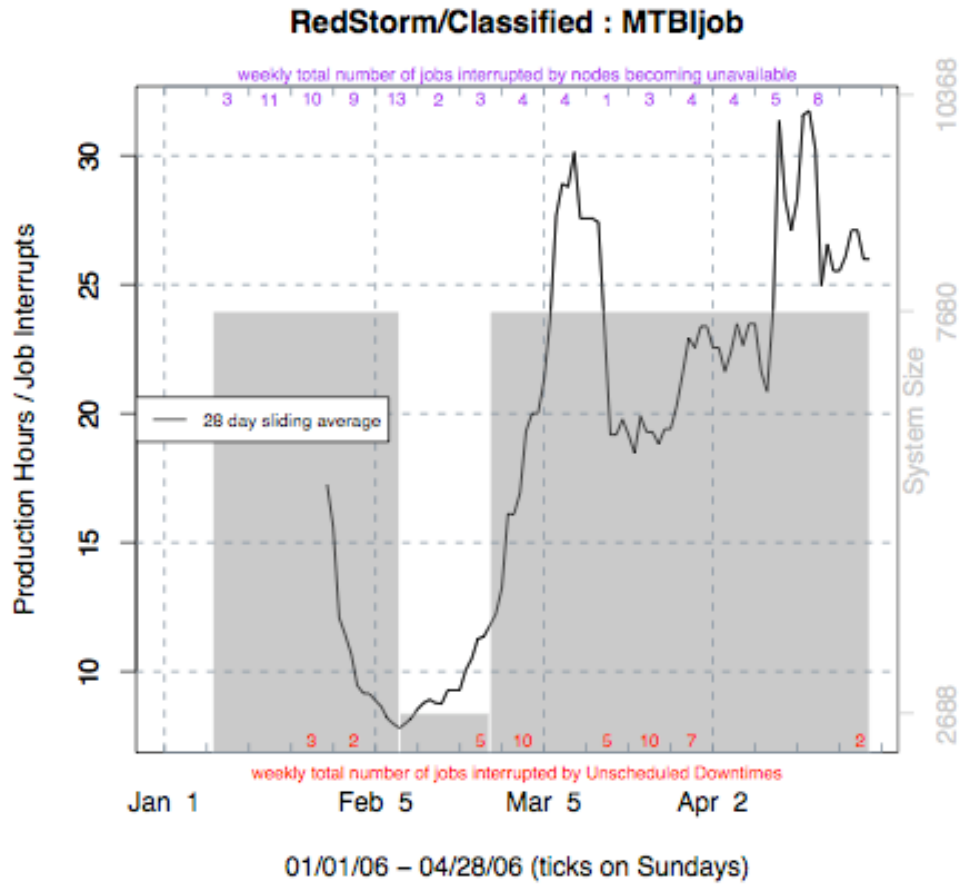
Direct hits cause job hangs or failures



Router failures also cause holes in the mesh, and so may induce job failures

MTBIjob


RedStorm/Unclassified : MTBIjob
weekly total number of jobs interrupted by node Unscheduled Downtimes
1 1 1 1 2 6 6 17 3 1 3 2
Production Hours / Job Interrupts
150
100
50
28 day sliding average
System Size
10368
7680
2688
4 38 5 3 10 11
weekly total number of jobs interrupted by system Unscheduled Downtimes
Jan 1
Feb 5
Mar 5
Apr 2
01/01/06 – 04/28/06 (ticks on Sundays)


RedStorm/Classified : MTBIjob
weekly total number of jobs interrupted by nodes becoming unavailable
3 11 10 9 13 2 3 4 4 1 3 4 4 5 8
Production Hours / Job Interrupts
30
25
20
15
10
28 day sliding average
System Size
10368
7680
2688
3 2 5 10 5 10 7 2
weekly total number of jobs interrupted by Unscheduled Downtimes
Jan 1
Feb 5
Mar 5
Apr 2
01/01/06 – 04/28/06 (ticks on Sundays)

Sandia National Laboratories

Lessons Learned

- Clear RAS definitions (states, failures, interrupts, …) need to be established (the HPC community should do this)!
- Means to easily capture data regarding the above is needed (e.g. vendors should supply this)!
- Need for the above is increasing with system size and complexity!
- Distilling RAS information into clear plots is challenging.
- Buy an XT3 today! ;)

RASM Toolkit

Input

(ASCII files)

Output

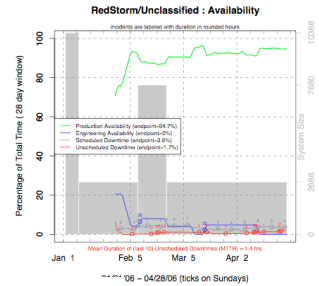
pdf|ps|png|…

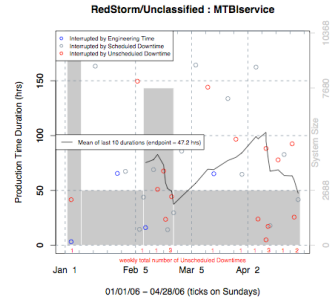
“state_transitions.tab”

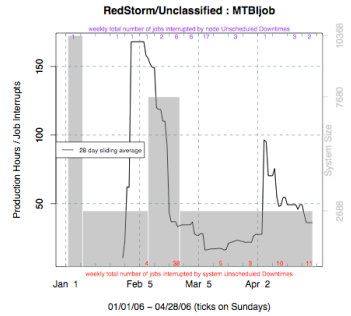
time state size

“job_interrupts.tab”

time interrupt_type







Written in R, includes documentation and example scripts.

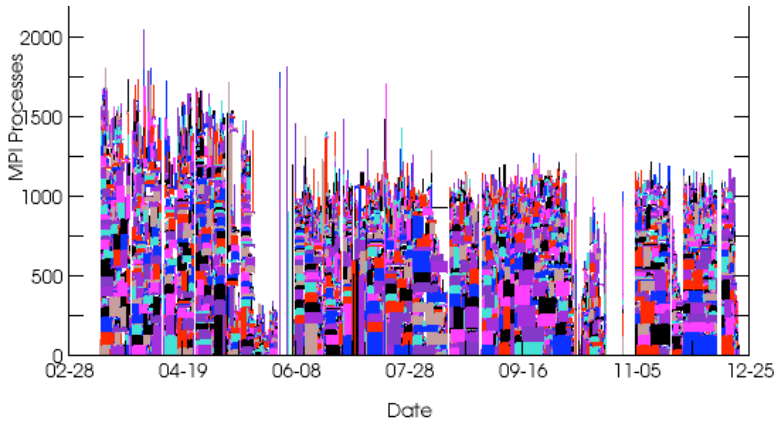
Nodehour-Scaled Metrics (TBD…)

Utilize workload information in order to measure work per interrupt rather than simply time per interrupt.

$$\text{MNBIservice}\,(\text{hours}) = \sum \frac{\text{Productive Nodehours}}{\text{Service Interrupts}}$$

$$\text{MNBIjob}\,(\text{hours}) = \sum \frac{\text{Productive Nodehours}}{\text{Job Interrupts}}$$

…

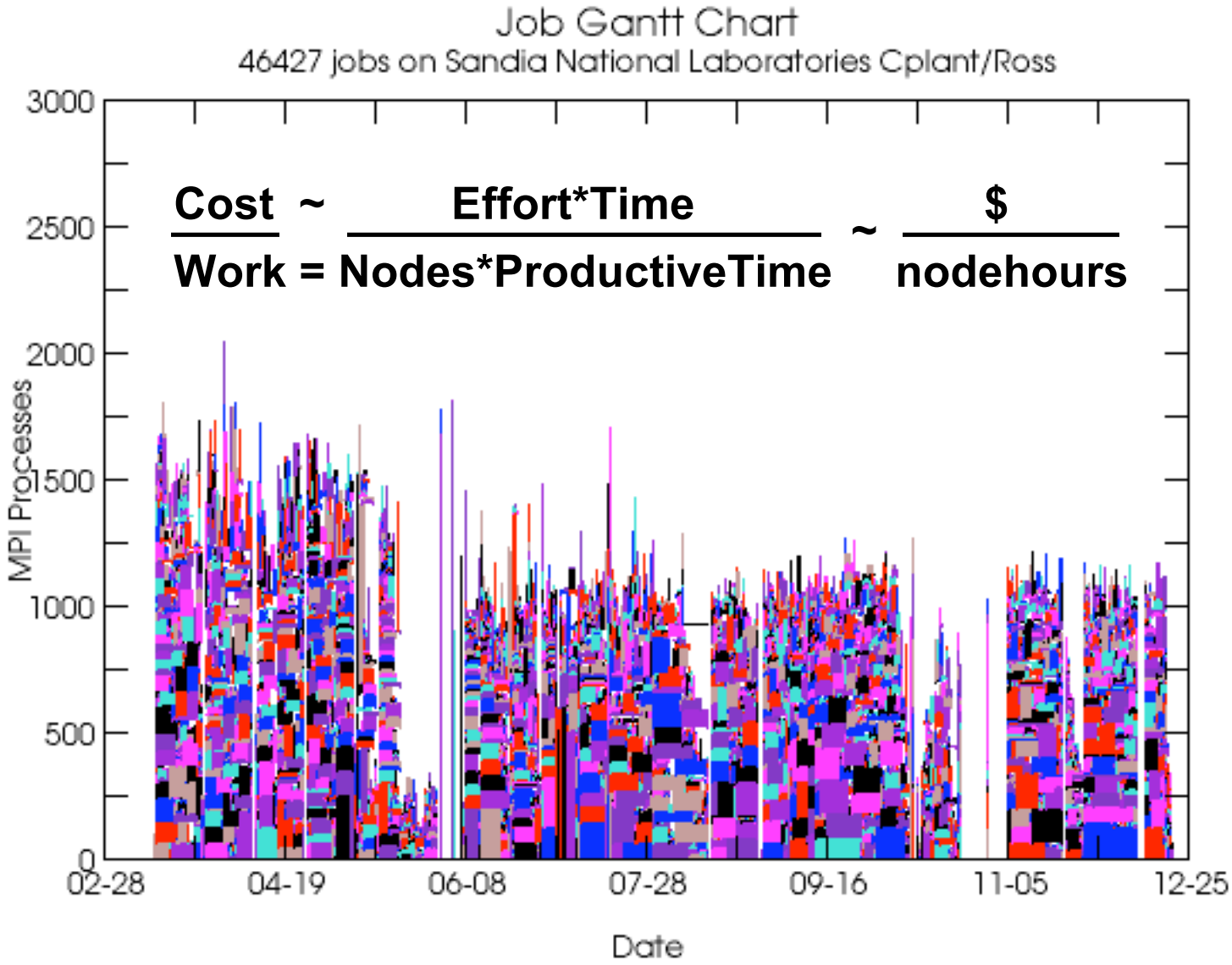


For More Info…

Sandia National Laboratories

See http://www.cs.sandia.gov/~jrstear/ras/

Workload information is vital!



$$\frac{\text{Cost}}{\text{Work = Nodes*ProductiveTime}} \sim \frac{\text{Effort*Time}}{} \sim \frac{\$}{\text{nodehours}}$$