Reliability Issues on ASCI Bluemountain and ASCI Q

Laura Davey

Andy Martinez

Georgia Pedicini

Phil Salazar

Background

- Accelerated Strategic Computing Initiative
- SGI Bluemountain; HP (Compaq) Q
- Stability

Game Plan

- Types of problems
- Metrics collected and reported
- Locally developed tools

Stability

- RAS
  - Reliability, Availability, Serviceability
- General theory of testing:
  - Unit, Integration, System

Problems - Unit

- Hardware
  - CPU
  - Memory
  - Power supplies
  - …
- Network
  - Interface

Problems - Integration

- Network topology
- File systems
- Operating system

Common thread -
difficulties scaling up to resource limit

Problems - System

- Operating System
- Scheduling
- MPI
- Diagnostic tools: monitoring, measurement, notification, optimization

Common threads -

scaling; capability implementation

RAS Data Collection

- When is Data Collected?
- What Data is Collected?
- Collection Process

REMEDY H/W & S/W MAINTENANCE SHEET

System:_____ SSE:_____ Operator:_____

Date:_____ Time Down:_____ Time Up:_____

Failure Message:_____

	Scheduled	Unscheduled

Reason:

Facilities:	Chillers	Environment	Power Outage	Power Spike	UPS

Hardware:					**Software**
Network	**Graphics**	**Fibre**	**System**	**Other**	
1 GBit Ethernet Card	Cable	Cable	CD ROM	Other	Alpha Server
1 MBit Ethernet Card	Fan Assembly	Controller	Console Interface Module	Undetermined	Cluster File System
Console Multi-port Module	GE Board	Drive	CPU	Unresolvable	Compilers and libraries
Ethernet Switch	Graphics Video Card	Fan Assembly	Fan Assembly		Disk IO, firmware and storage
Flash Card	Midplane	Fibre Channel Port Adptr	Fibre XIO		DST-Flare Code
Gig E Connection	Power Supply	Fibre Disk Enclosure	Floppy Diskette Drive		DST-Scan for errors/scratch
Hippi Interface		Fibre Switch	GS Hierarchical Switch		DST-Upgrd/Instl 3rd party Sftwr
128 Port HL Card		Fibre XIO	Hippi XIO		DST-Upgrd/vendor Sftwr
16 Port Switch		LCC Card	Memory DIMM		HPSS
Fan Assembly		Midplane	Midplane		HIPPI-800
Switch Back Plane		PCI Fibre Channel Adptr	MMSC		LSF
Switch Clock Card		Power Supply	MSC		Modify system config files
Switch Interconnect Cable			Node Board		MPI
Terminal Server			PCI Back Plane		Netrwork
			PCI IO Module		Other Software
			Power Supply		Parallel File System
			Bach Plane Assembly		Patch Install
			Clock Distribution Module		Resource Mgmt System
			Clock Splitter Module		Security Software
			Directory Module		Tru 64
			Global Port Module		Upgrade/Install 3rd Party Sftw
			ROM BIOS Battery		Upgrade/vendor Sftw
			SCSI Adapter Card		User code
			SCSI Cable		User tools (debuggers…)
			SCSI Controller		
			SCSI Drive		

Part Replaced:_____

Part Location:_____

Comments:_____

Verification of RAS Data

- Data is checked with other reports
- Vendor representative meetings
- Changes incorporated back into database

RAS Reporting Process

- Who's interested?
- Published Reports
  - **http://icnn → <clusterName>Today → System Reliability/Availability**
- Types of reports and importance

MTTF per Month in Hours											
	May	June	July	August	September	October	November	December	January	February	March
	17.7	16.7	24	20.1	17.1	24.8	15.7	16.2	12.6	16.8	21.9

Categorized Failures per Month											
	May	June	July	August	September	October	November	December	January	February	March
Hardware	30	32	23	24	36	28	38	35	50	32	26
Other	12	11	8	13	6	2	8	11	9	8	8
Total	42	43	31	37	42	30	46	46	59	40	34

MTTI Scheduled / Unscheduled in Hours											
	May	June	July	August	September	October	November	December	January	February	March
Scheduled	82.7	32.7	57.2	67.6	72	93.1	51.4	74.4	67.6	56	106.3
Unschedule	17.7	16.7	24	20.1	17.1	24.8	15.7	16.2	12.6	16.8	21.9
Combined	14.6	11.1	16.9	15.5	13.8	19.6	12	13.3	10.6	12.9	18.2

Number of Interrupts Scheduled / Unscheduled											
	May	June	July	August	September	October	November	December	January	February	March
Scheduled	9	22	13	11	10	8	14	10	11	12	7
Unschedule	42	43	31	37	42	30	46	46	59	40	34
Total	51	65	44	48	52	38	60	56	70	52	41

Monthly

MTTF per Week in Hours										
	Wk1	Wk2	Wk3	Wk4	Wk5	Wk6	Wk7	Wk8	Wk9	Wk10
	3.7	3.4	2.1	4.8	7.3	5.8	3.2	3.2	21	21

Categorized Failures per Week										
	Wk1	Wk2	Wk3	Wk4	Wk5	Wk6	Wk7	Wk8	Wk9	Wk10
Hardware	36	39	64	16	23	21	46	45	7	5
Other	9	10	15	19	0	8	6	8	1	3
Total	45	49	79	35	23	29	52	53	8	8

Number of Interrupts scheduled / unscheduled per Week										
	Wk1	Wk2	Wk3	Wk4	Wk5	Wk6	Wk7	Wk8	Wk9	Wk10
Scheduled	5	1	1	0	1	1	2	1	1	2
Unschedule	45	49	79	35	23	29	52	53	8	8
Total	50	50	80	35	24	30	54	54	9	10

MTTI Scheduled / Unscheduled per Week in Hours										
	Wk1	Wk2	Wk3	Wk4	Wk5	Wk6	Wk7	Wk8	Wk9	Wk10
Scheduled	33.6	126	126	168	126	125.3	84	126	126	84
Unschedule	3.7	3.4	2.1	4.8	7.3	5.8	3.2	3.2	21	21
Combined	3.4	3.4	2.1	4.8	7	5.6	3.1	3.1	18.7	16.8

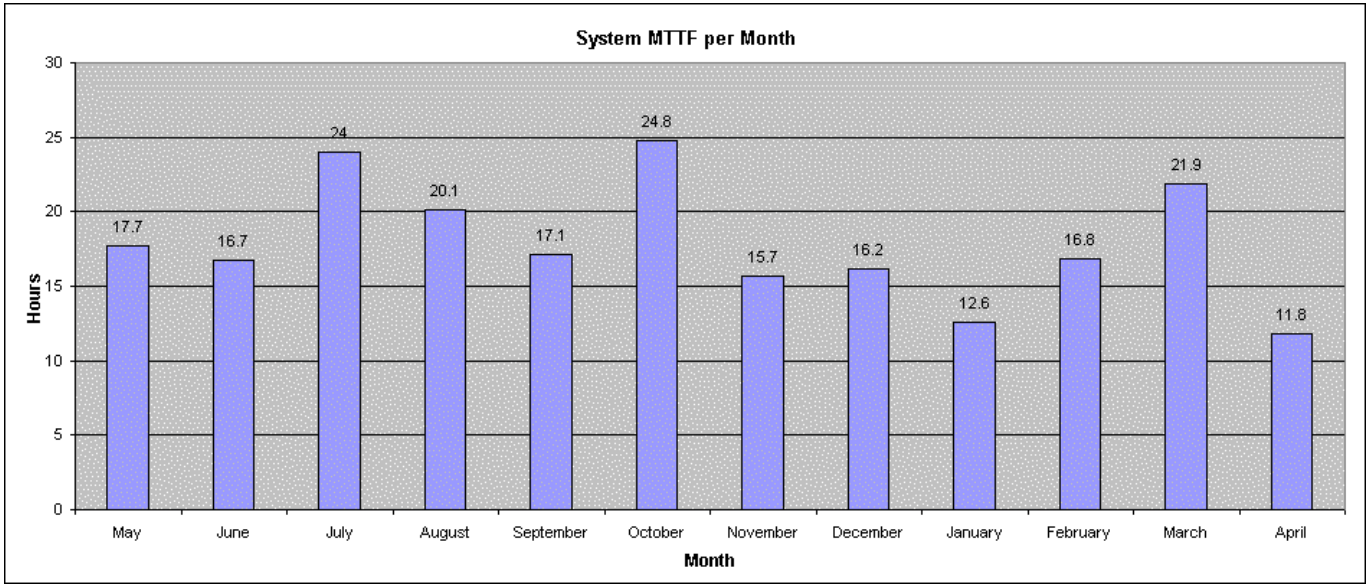
Weekly

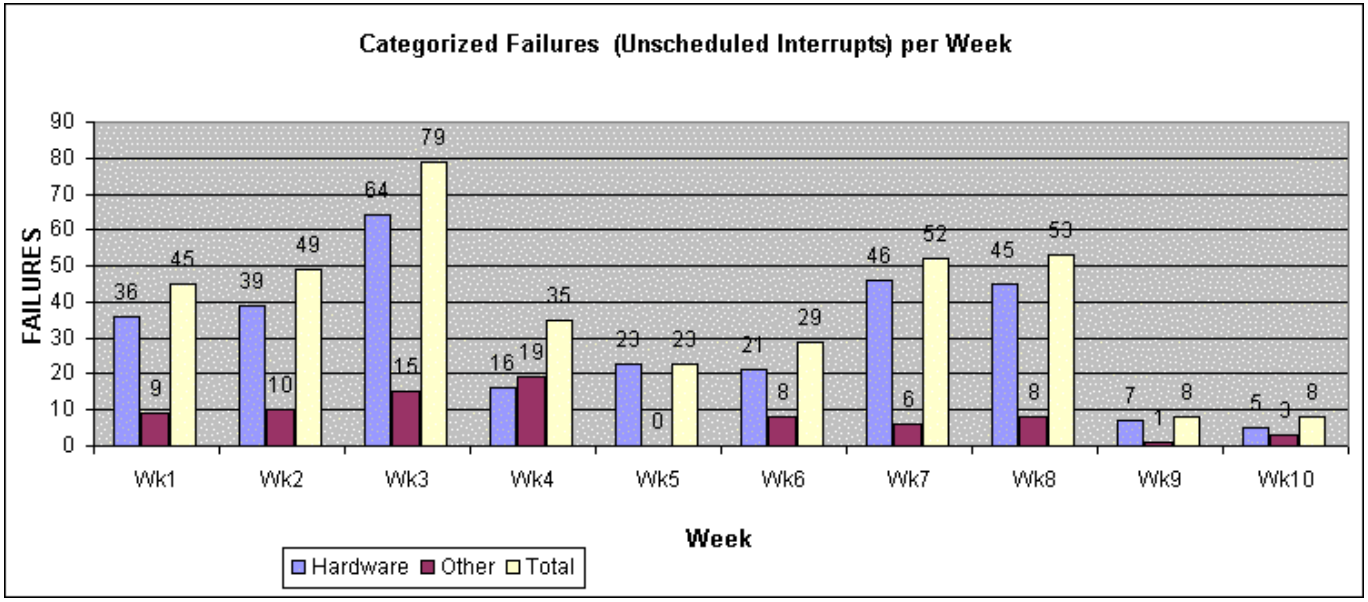
Los Alamos NATIONAL LABORATORY

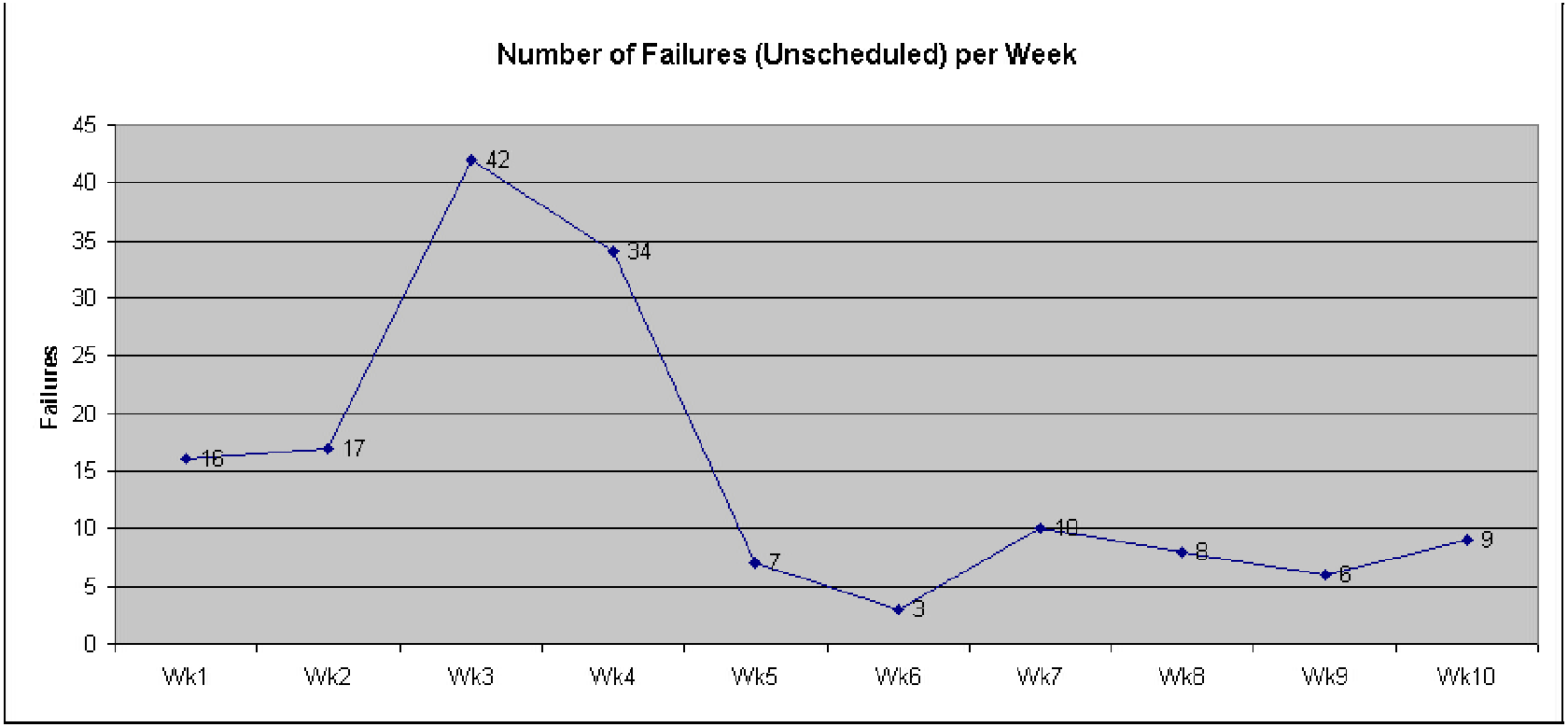
ASCI Accelerated Strategic Computing Initiative Los Alamos Sandia Lawrence Livermore

Detailed ***Scheduled*** and ***Unscheduled*** Categorized Hardware Interrupts

	May	June	July	August	September	October	November	December	January	February	March	April
Attenuators												
Compressor												
CPU		1					1	1				
Cray Disk												
Cray Link Cable					1							
DG Board												
Disk Cabinet		1										
Fan Assembly			1			1		2	2	1		
Fibre Cable							1					1
Fibre Channel Port adaptor												1
Fibre Drive		1										
Fibre Raid Controller	2	1			1		1			2		
Fibre Raid Fan Assembly												
Fibre Raid LCC Card												
Fibre Raid Midplane												
Fibre Raid Power Supply												
Fibre XIO		1	2									
Fddi												
GE Board										1		
Gig E Connection									1	1		
GigEXIO								1				
Gig E Switch		1			1				15	1	2	
Graphic-IR pipes		1		1	1				2			
GSN								2			3	
HEU												
Hippi Interface	1											2
Hippi Switch	1	2										
Hippi Xio	2	11	5	6	4	2	6		3	2	2	8
Hisp												
HVDC												
IOS Buffer Memory												
IOS CPU												
IOS Disk Logic												
IO6												
Ktown Board	1								1			
Maintenance	2	3	2	2	2	2	3		2	3	3	2
Motor Generator												
MIA	1									1		
Memory Dimm	11	18	12	12	15	15	17	12	14	12	7	24
Memory Module												
Mid-plane	1		1	1		2	8	5	2	1	1	2
Module Assembly												
MSC Board									1			
Node Board	7	5	5	6	11	8	6	11	8	13	8	14
Other	4											
PCI Shoebox												
Power Supply	3	4	2	1	6	2	7	7	7	2	4	6
RM Board												
Router Board				1						1		
SCSI Cable												
SCSI Controller												
SCSI Drive	1											1
SCSI Playdough Box												
SSD Logic												
SSD Memory Module												
Support Chassis												
System Controller												
Vhisp												
WACS Logic												
Xtown Board												
Total	37	50	30	30	42	32	50	41	58	41	30	61


System MTTF per Month
Hours
30
25
20
15
10
5
0
17.7
16.7
24
20.1
17.1
24.8
15.7
16.2
12.6
16.8
21.9
11.8
May
June
July
August
September
October
November
December
January
February
March
April
Month


Categorized Failures (Unscheduled Interrupts) per Week
FAILURES
90
80
70
60
50
40
30
20
10
0
36
9
45
39
10
49
64
15
79
16
19
35
23
0
23
21
8
29
46
6
52
45
8
53
7
1
8
5
3
8
Wk1
Wk2
Wk3
Wk4
Wk5
Wk6
Wk7
Wk8
Wk9
Wk10
Week
Hardware
Other
Total



Los Alamos NATIONAL LABORATORY

ASCI Accelerated Strategic Computing Initiative

Local Test Tool Features

- Framework -nested set of scripts
- Flexible, extensible
- Parameterized
- Consolidated pass/fail report for test series

What are we checking?

- System software unit testing
- …followed by crossbox communications…
- …followed by sample crossbox message-passing routines, and “typical user” tasks…
- …culminating in an industry benchmark

Success / Failure

- Pass? Go on to next
- Success means ...

- Fail? Examine reason.
- May be an acceptable aberration
- May require correction (HW or SW)
- May require back-tracking tests

Conclusions

Combined approach allows us to

- Isolate …
- Characterize …
- Focus attention on …
  - a variety of problems