SAND2009-6703C

Data Fusion and Statistical Analysis: Piercing the Darkness of the Black Box

Jim Brandt, Frank Chen, Vince De Sapio, Ann Gentile, Jackson Mayo, Phillippe Pebay, Diana Roe, David Thompson, Matthew Wong

With support from:

Marcus Epperson and Jerry Smith

This work supported by the ASC Computational Systems and Software Environments (CSSE) program element

Black Box Defn.

www.thefreedictionary.com

a. A device or theoretical construct with known or specified performance characteristics but unknown or unspecified constituents and means of operation.

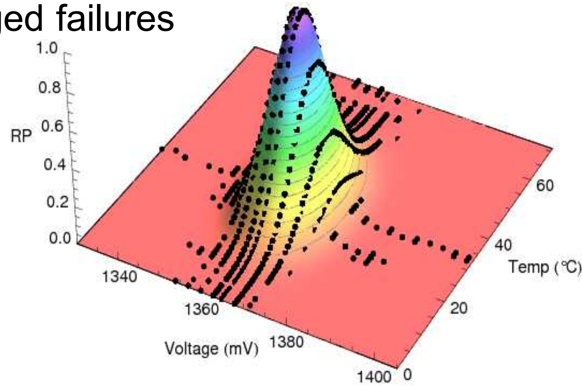
b. Something that is mysterious, especially as to function.

Motivation

- Large amount of time spent/(wasted?) in checkpointing
- Black Box approach to failures in platforms implies scaling wall
  - Best we can do is adjust checkpointing freq based on the measured system average MTTI
- Assumption -- If we understood failure mechanisms
  - Instrument to get precursor warnings where possible
  - System and Application interfaces that would facilitate such recovery
  - Know where to invest more up front for more robust systems
    - Power supply redundancy cost/fault tolerance tradeoff
    - Memory socket cost/MTTF tradeoff

Resilience Fault Prediction Strategy

- Discover predictors, accuracy, time windows, and coverage with respect to all non-recoverable faults
  - Scalable data collection
    - HW related metrics
      - Limited by current instrumentation
      - Discovery can help drive future system instrumentation
    - System related metrics
      - RM databases, log files, troubleshooting notes, etc.
      - Work with System Administrators to capture as much as possible
    - Not available
      - Human errors, power grid outages, etc.
  - Scalable data analysis
    - Definition of analysis methods that make sense given the data and time scales
      - Currently: correlate low probability behaviors with logged failures
    - Efficient data exploration tools
      - UI
      - Visualization
- Quantify Prediction Effectiveness



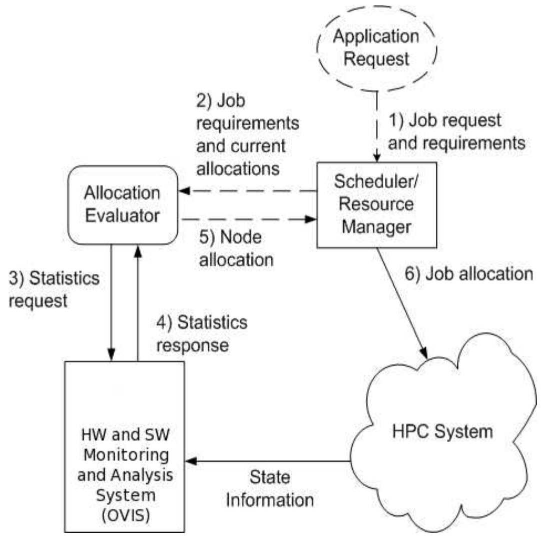
Approaches

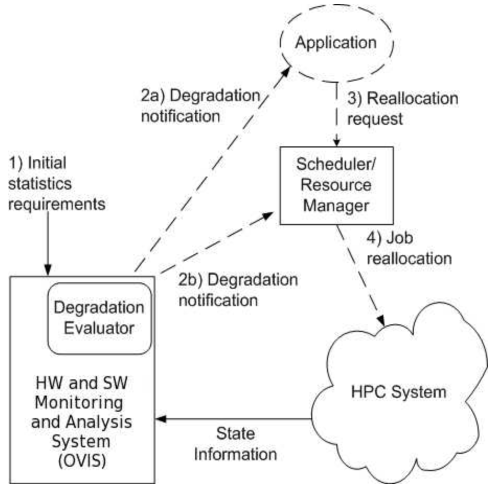
Seek "real time" functional component (Memory bank, core, communication bus, power supply sub-systems, etc.) level understanding of health

- Statistical analysis
  - Learn statistical characteristics of components that lead to both normal and abnormal operation of a system or platform
- Data Fusion
  - Look for correlations in disparate data to facilitate understanding of component, system and application level behavioral interaction, operation, and failure
    - Could be outlier or smack in the middle of normal operational regions -- TBD

Ultimate Goal

- Health based allocation given parameters of application run
- Preemptive task/state migration based on failure prediction
- Targeted checkpoint frequency based on health (e.g. resource with higher chance of failure might save state frequently but this would be much faster than saving the aggregate. Have to address how much larger the memory footprint gets for healthy resources having to maintain state in case of rollback)


Application Request
1) Job request and requirements
2) Job requirements and current allocations
Allocation Evaluator
Scheduler/ Resource Manager
5) Node allocation
3) Statistics request
4) Statistics response
6) Job allocation
HW and SW Monitoring and Analysis System (OVIS)
HPC System
State Information

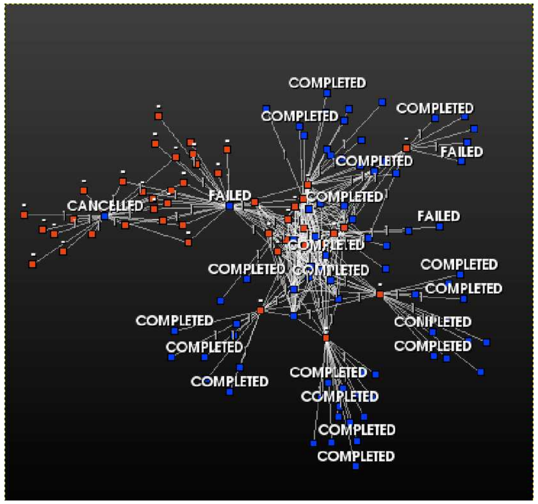

Application
2a) Degradation notification
3) Reallocation request
1) Initial statistics requirements
Scheduler/ Resource Manager
4) Job reallocation
2b) Degradation notification
Degradation Evaluator
HW and SW Monitoring and Analysis System (OVIS)
HPC System
State Information

How does this differ?

- Many system level studies (Black Box approach) give MTTI over the system that can lead to extrapolated MTTI for a pool size – Same studies show this approach to checkpoint frequency calculation doesn’t scale

Statistical Approaches

- Descriptive
  - Numeric metrics such as mem err rates, cpu utlization, voltage, PS ripple
- Correlation
  - Looking at numeric ensemble behaviors in run-time environment
- Time Series
  - Taking into account temporal displacements
  - Temporal ensemble behaviors relating to actual codes



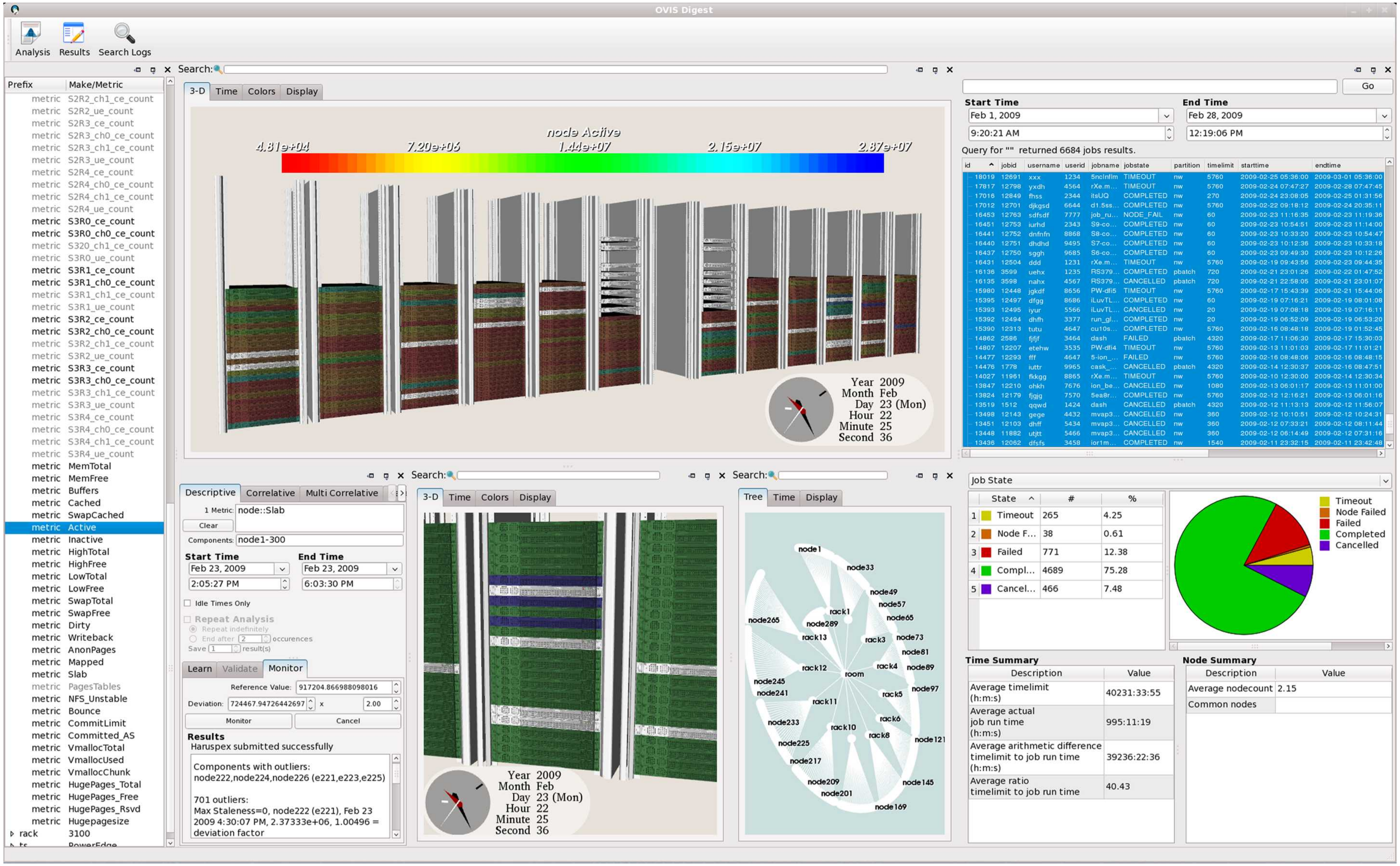
- Graph based
  - Exploration of both numeric and text (log) data for telltale relationships using clustering techniques
- Incorporate both parametric and failure data to produce models that target the prediction of those failures
  - Looking for numeric metric correlations with failure that aren't tied to anomalous behavior
  - Treats failure as a generic event without subtypes

Data Fusion

Cockpit approach with interacting capabilities

- Perform analyses on aggregate of parametric data, Resource Manager (Slurm) data, Syslog, and Console logs
    - Relate user, application (problems here), resources used, application and resource based failures

Data Fusion and Interacting Capabilities



Prefix	Make/Metric
metric	S2R2_ch1_ce_count
metric	S2R2_ue_count
metric	S2R3_ce_count
metric	S2R3_ch0_ce_count
metric	S2R3_ch1_ce_count
metric	S2R3_ue_count
metric	S2R4_ce_count
metric	S2R4_ch0_ce_count
metric	S2R4_ch1_ce_count
metric	S2R4_ue_count
metric	S3R0_ce_count
metric	S3R0_ch0_ce_count
metric	S320_ch1_ce_count
metric	S3R0_ue_count
metric	S3R1_ce_count
metric	S3R1_ch0_ce_count
metric	S3R1_ch1_ce_count
metric	S3R1_ue_count
metric	S3R2_ce_count
metric	S3R2_ch0_ce_count
metric	S3R2_ch1_ce_count
metric	S3R2_ue_count
metric	S3R3_ce_count
metric	S3R3_ch0_ce_count
metric	S3R3_ch1_ce_count
metric	S3R3_ue_count
metric	S3R4_ce_count
metric	S3R4_ch0_ce_count
metric	S3R4_ch1_ce_count
metric	S3R4_ue_count
metric	MemTotal
metric	MemFree
metric	Buffers
metric	Cached
metric	SwapCached
metric	Active
metric	Inactive
metric	HighTotal
metric	HighFree
metric	LowTotal
metric	LowFree
metric	SwapTotal
metric	SwapFree
metric	Dirty
metric	Writeback
metric	AnonPages
metric	Mapped
metric	Slab
metric	PagesTables
metric	NFS_Unstable
metric	Bounce
metric	CommitLimit
metric	Committed_AS
metric	VmallocTotal
metric	VmallocUsed
metric	VmallocChunk
metric	HugePages_Total
metric	HugePages_Free
metric	HugePages_Rsvd
metric	Hugepagesize
rack	3100
ts	PowerEdge

Start Time Feb 1, 2009 9:20:21 AM
End Time Feb 28, 2009 12:19:06 PM

Query for "" returned 6684 jobs results.

id	jobid	username	userid	jobname	jobstate	partition	timelimit	starttime	endtime
18019	12691	xxx	1234	5nclnflm	TIMEOUT	nw	5760	2009-02-25 05:36:00	2009-03-01 05:36:00
17817	12798	yxdh	4564	rXe.m...	TIMEOUT	nw	5760	2009-02-24 07:47:27	2009-02-28 07:47:45
17016	12849	fhss	2344	itsUQ	COMPLETED	nw	270	2009-02-24 23:08:05	2009-02-25 01:31:56
17012	12701	djkgsd	6644	d1.5ss...	COMPLETED	nw	5760	2009-02-22 09:18:12	2009-02-24 20:35:11
16453	12763	sdfsdf	7777	job_ru...	NODE_FAIL	nw	60	2009-02-23 11:16:35	2009-02-23 11:19:36
16451	12753	iurhd	2343	S9-co...	COMPLETED	nw	60	2009-02-23 10:54:51	2009-02-23 11:14:00
16441	12752	dnfnfn	8868	S8-co...	COMPLETED	nw	60	2009-02-23 10:33:20	2009-02-23 10:54:47
16440	12751	dhdhd	9495	S7-co...	COMPLETED	nw	60	2009-02-23 10:12:36	2009-02-23 10:33:18
16437	12750	sggh	9685	S6-co...	COMPLETED	nw	60	2009-02-23 09:49:30	2009-02-23 10:12:26
16431	12504	ddd	1231	rXe.m...	TIMEOUT	nw	5760	2009-02-19 09:43:56	2009-02-23 09:44:35
16136	3599	uehx	1235	RS379...	COMPLETED	pbatch	720	2009-02-21 23:01:26	2009-02-22 01:47:52
16135	3598	nahx	4567	RS379...	CANCELLED	pbatch	720	2009-02-21 22:58:05	2009-02-21 23:01:07
15980	12448	jgkdf	8656	PW-dfi5	TIMEOUT	nw	5760	2009-02-17 15:43:39	2009-02-21 15:44:06
15395	12497	dfgg	8686	iLuvTL...	COMPLETED	nw	60	2009-02-19 07:16:21	2009-02-19 08:01:08
15393	12495	iyur	5566	iLuvTL...	CANCELLED	nw	20	2009-02-19 07:08:18	2009-02-19 07:16:11
15392	12494	dhfh	3377	run_gl...	COMPLETED	nw	20	2009-02-19 06:52:09	2009-02-19 06:53:20
15390	12313	tutu	4647	cu10s...	COMPLETED	nw	5760	2009-02-16 08:48:18	2009-02-19 01:52:45
14862	2586	fjfjf	3464	dash	FAILED	pbatch	4320	2009-02-17 11:06:30	2009-02-17 15:30:03
14807	12207	etehw	3535	PW-dfi4	TIMEOUT	nw	5760	2009-02-13 11:01:03	2009-02-17 11:01:21
14477	12293	fff	4647	5-ion_...	FAILED	nw	5760	2009-02-16 08:48:06	2009-02-16 08:48:15
14476	1778	iuttr	9965	cask_...	CANCELLED	pbatch	4320	2009-02-14 12:30:37	2009-02-16 08:47:51
14027	11961	fkkgg	8865	rXe.m...	TIMEOUT	nw	5760	2009-02-10 12:30:00	2009-02-14 12:30:34
13847	12210	ohkh	7676	ion_be...	CANCELLED	nw	1080	2009-02-13 06:01:17	2009-02-13 11:01:00
13824	12179	fjgjg	7570	5ea8r...	COMPLETED	nw	5760	2009-02-12 12:16:21	2009-02-13 06:01:16
13519	1512	qqwd	1424	dash	CANCELLED	pbatch	4320	2009-02-12 11:13:13	2009-02-12 11:56:07
13498	12143	gege	4432	mvap3...	CANCELLED	nw	360	2009-02-12 10:10:51	2009-02-12 10:24:31
13451	12103	dhff	5434	mvap3...	CANCELLED	nw	360	2009-02-12 07:33:21	2009-02-12 08:11:44
13448	11882	utjtt	5466	mvap3...	CANCELLED	nw	360	2009-02-12 06:14:49	2009-02-12 07:31:16
13436	12062	dfsfs	3458	ior1m...	COMPLETED	nw	1540	2009-02-11 23:32:15	2009-02-11 23:42:48

Descriptive | Correlative | Multi Correlative

1 Metric: node::Slab
Clear
Components: node1-300

Start Time Feb 23, 2009 2:05:27 PM
End Time Feb 23, 2009 6:03:30 PM

☐ Idle Times Only

☐ Repeat Analysis
● Repeat indefinitely
○ End after 2 occurences
Save 1 result(s)

Learn | Validate | Monitor

Reference Value: 917204.866988098016
Deviation: 724467.94726442697 x 2.00
Monitor | Cancel

Results
Haruspex submitted successfully

Components with outliers:
node222,node224,node226 (e221,e223,e225)

701 outliers:
Max Staleness=0, node222 (e221), Feb 23 2009 4:30:07 PM, 2.37333e+06, 1.00496 = deviation factor

Job State

	State	#	%
1	🟨 Timeout	265	4.25
2	🟧 Node F...	38	0.61
3	🟥 Failed	771	12.38
4	🟩 Compl...	4689	75.28
5	🟪 Cancel...	466	7.48

🟨 Timeout
🟧 Node Failed
🟥 Failed
🟩 Completed
🟪 Cancelled

Time Summary

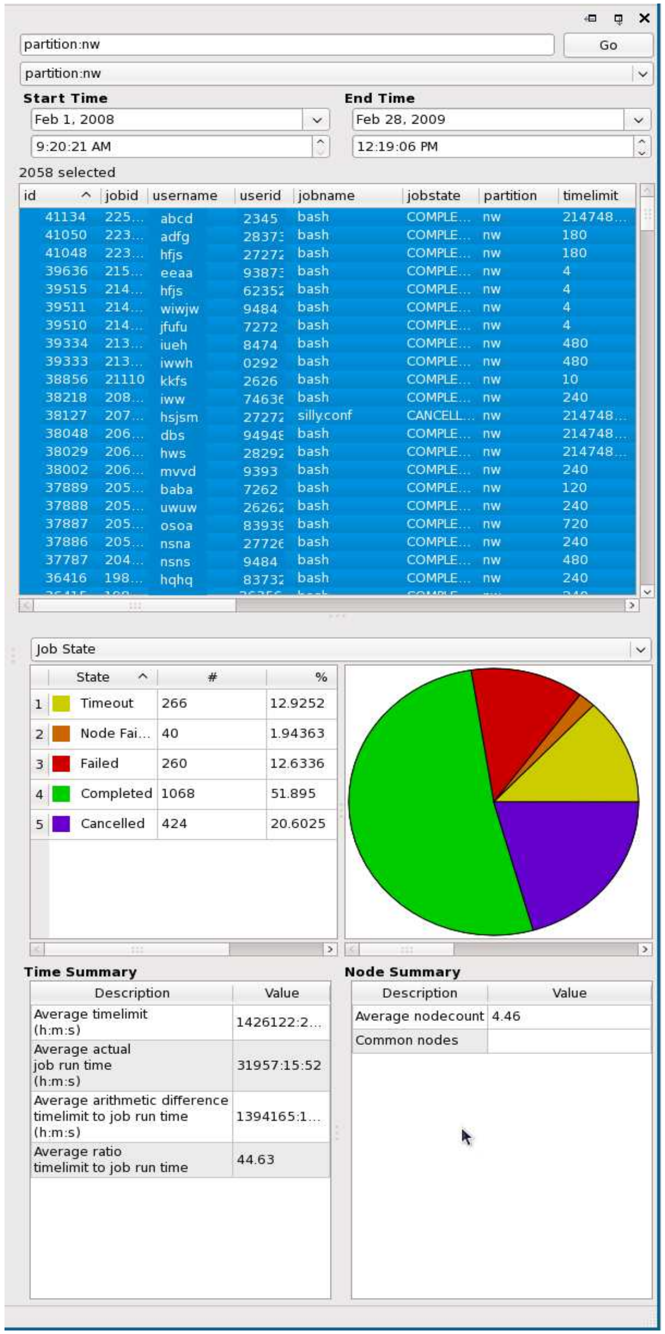
Description	Value
Average timelimit (h:m:s)	40231:33:55
Average actual job run time (h:m:s)	995:11:19
Average arithmetic difference timelimit to job run time (h:m:s)	39236:22:36
Average ratio timelimit to job run time	40.43

Node Summary

Description	Value
Average nodecount	2.15
Common nodes	

Scheduler and Log Search

- Integrated Scheduler and Error Log Searches
- Searchable:
  - e.g., OOM && node 234
- Aggregate statistics
- Pie subselection
  - e.g. Distribution of failed jobs by user



partition:nw

Go

partition:nw

Start Time

Feb 1, 2008

9:20:21 AM

End Time

Feb 28, 2009

12:19:06 PM

2058 selected

id	jobid	username	userid	jobname	jobstate	partition	timelimit
41134	225...	abcd	2345	bash	COMPLE...	nw	214748...
41050	223...	adfg	28373	bash	COMPLE...	nw	180
41048	223...	hfjs	27272	bash	COMPLE...	nw	180
39636	215...	eeaa	93873	bash	COMPLE...	nw	4
39515	214...	hfjs	62352	bash	COMPLE...	nw	4
39511	214...	wiwjw	9484	bash	COMPLE...	nw	4
39510	214...	jfufu	7272	bash	COMPLE...	nw	4
39334	213...	iueh	8474	bash	COMPLE...	nw	480
39333	213...	iwwh	0292	bash	COMPLE...	nw	480
38856	21110	kkfs	2626	bash	COMPLE...	nw	10
38218	208...	iww	74636	bash	COMPLE...	nw	240
38127	207...	hsjsm	27272	silly.conf	CANCELL...	nw	214748...
38048	206...	dbs	94948	bash	COMPLE...	nw	214748...
38029	206...	hws	28292	bash	COMPLE...	nw	214748...
38002	206...	mvvd	9393	bash	COMPLE...	nw	240
37889	205...	baba	7262	bash	COMPLE...	nw	120
37888	205...	uwuw	26262	bash	COMPLE...	nw	240
37887	205...	osoa	83939	bash	COMPLE...	nw	720
37886	205...	nsna	27726	bash	COMPLE...	nw	240
37787	204...	nsns	9484	bash	COMPLE...	nw	480
36416	198...	hqhq	83732	bash	COMPLE...	nw	240

Job State

	State	#	%
1	🟨 Timeout	266	12.9252
2	🟧 Node Fai...	40	1.94363
3	🟥 Failed	260	12.6336
4	🟩 Completed	1068	51.895
5	🟪 Cancelled	424	20.6025

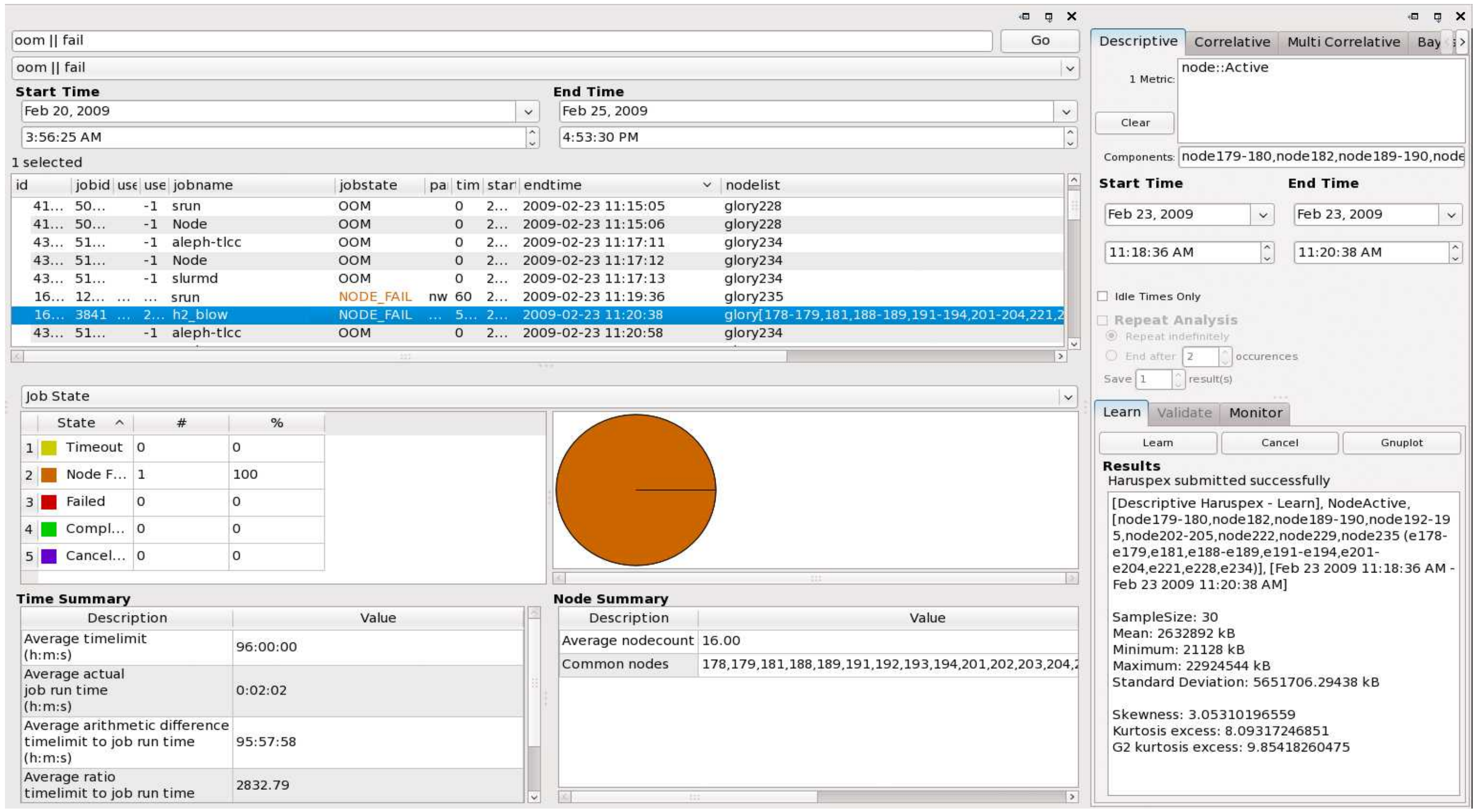
Time Summary

Description	Value
Average timelimit (h:m:s)	1426122:2...
Average actual job run time (h:m:s)	31957:15:52
Average arithmetic difference timelimit to job run time (h:m:s)	1394165:1...
Average ratio timelimit to job run time	44.63

Node Summary

Description	Value
Average nodecount	4.46
Common nodes	

Scheduler and Log data Drive Analysis



oom || fail

Start Time: Feb 20, 2009 3:56:25 AM

End Time: Feb 25, 2009 4:53:30 PM

1 selected

id	jobid	use	use	jobname	jobstate	pa	tim	star	endtime	nodelist
41...	50...		-1	srun	OOM		0	2...	2009-02-23 11:15:05	glory228
41...	50...		-1	Node	OOM		0	2...	2009-02-23 11:15:06	glory228
43...	51...		-1	aleph-tlcc	OOM		0	2...	2009-02-23 11:17:11	glory234
43...	51...		-1	Node	OOM		0	2...	2009-02-23 11:17:12	glory234
43...	51...		-1	slurmd	OOM		0	2...	2009-02-23 11:17:13	glory234
16...	12...	...	...	srun	NODE_FAIL	nw	60	2...	2009-02-23 11:19:36	glory235
16...	3841	...	2...	h2_blow	NODE_FAIL	...	5...	2...	2009-02-23 11:20:38	glory[178-179,181,188-189,191-194,201-204,221,2
43...	51...		-1	aleph-tlcc	OOM		0	2...	2009-02-23 11:20:58	glory234

Job State

	State	#	%
1	🟨 Timeout	0	0
2	🟧 Node F...	1	100
3	🟥 Failed	0	0
4	🟩 Compl...	0	0
5	🟪 Cancel...	0	0

Time Summary

Description	Value
Average timelimit (h:m:s)	96:00:00
Average actual job run time (h:m:s)	0:02:02
Average arithmetic difference timelimit to job run time (h:m:s)	95:57:58
Average ratio timelimit to job run time	2832.79

Node Summary

Description	Value
Average nodecount	16.00
Common nodes	178,179,181,188,189,191,192,193,194,201,202,203,204,2

Descriptive | Correlative | Multi Correlative | Bay

1 Metric: node::Active

Components: node179-180,node182,node189-190,node

Start Time Feb 23, 2009 11:18:36 AM

End Time Feb 23, 2009 11:20:38 AM

Idle Times Only

Repeat Analysis

Repeat indefinitely

End after 2 occurences

Save 1 result(s)

Learn | Validate | Monitor

Results

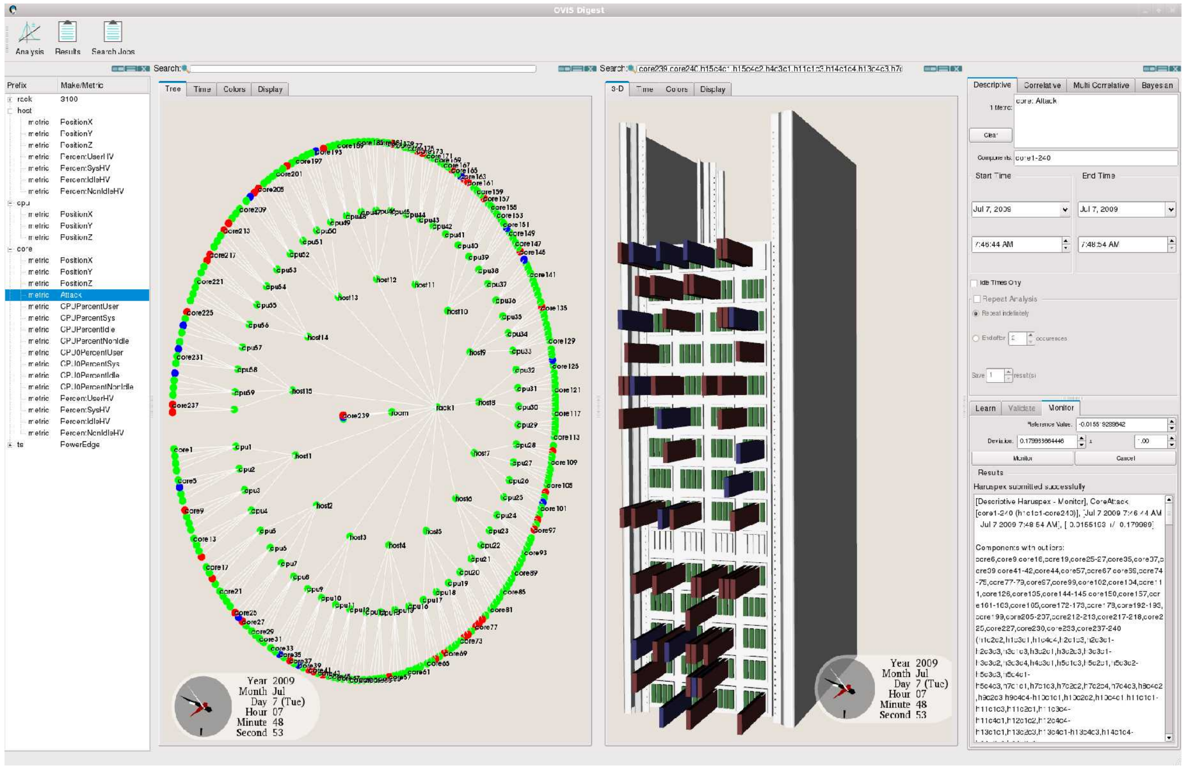
Haruspex submitted successfully

```
[Descriptive Haruspex - Learn], NodeActive,
[node179-180,node182,node189-190,node192-19
5,node202-205,node222,node229,node235 (e178-
e179,e181,e188-e189,e191-e194,e201-
e204,e221,e228,e234)], [Feb 23 2009 11:18:36 AM -
Feb 23 2009 11:20:38 AM]

SampleSize: 30
Mean: 2632892 kB
Minimum: 21128 kB
Maximum: 22924544 kB
Standard Deviation: 5651706.29438 kB

Skewness: 3.05310196559
Kurtosis excess: 8.09317246851
G2 kurtosis excess: 9.85418260475
```

- Time and component relationship of OOM killer and failed job leads to invocation of memory utilization analysis

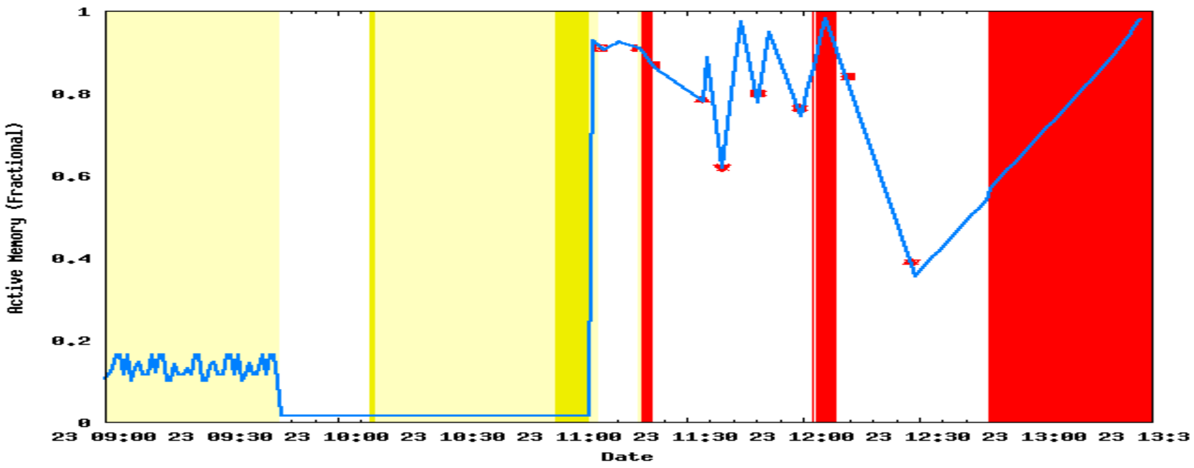
Data Fusion Visualizations

- Logical and Physical Displays
- Colored by analysis results



```
Reference Value: -0.0155 9289642
Deviation: 0.179983664446 x 1.00
Results
Haruspex submitted successfully
[Descriptive Haruspex - Monitor], CoreAttack
[core1-240 (h1c1c1-core240)], [Jul 7 2009 7:46:44 AM
 Jul 7 2009 7:48:54 AM], [ 0.0155193 +/- 0.179989]

Components with outliers:
core6,core9,core16,core19,core25-27,core35,core37,c
ore39,core41-42,core44,core57,core67,core69,core74
-75,core77-79,core97,core99,core102,core104,core11
1,core126,core135,core144-145,core150,core157,cor
e161-163,core165,core172-173,core178,core192-193,
core199,core205-207,core212-213,core217-218,core2
25,core227,core230,core233,core237-240
(h1c2c2,h1c3c1,h1c4c4,h2c1c3,h2c3c1-
h2c3c3,h3c1c3,h3c2c1,h3c2c3,h3c3c1-
h3c3c2,h3c3c4,h4c3c1,h5c1c3,h5c2c1,h5c3c2-
h5c3c3,h5c4c1-
h5c4c3,h7c1c1,h7c1c3,h7c2c2,h7c2c4,h7c4c3,h8c4c2
,h9c2c3,h9c4c4-h10c1c1,h10c2c2,h10c4c1,h11c1c1-
h11c1c3,h11c2c1,h11c3c4-
h11c4c1,h12c1c2,h12c4c4-
h13c1c1,h13c2c3,h13c4c1-h13c4c3,h14c1c4-
```

- Textual Job data – state and duration (shaded)
- Textual log data (x'es)
- Numerical data (blue)



Data Acquisition

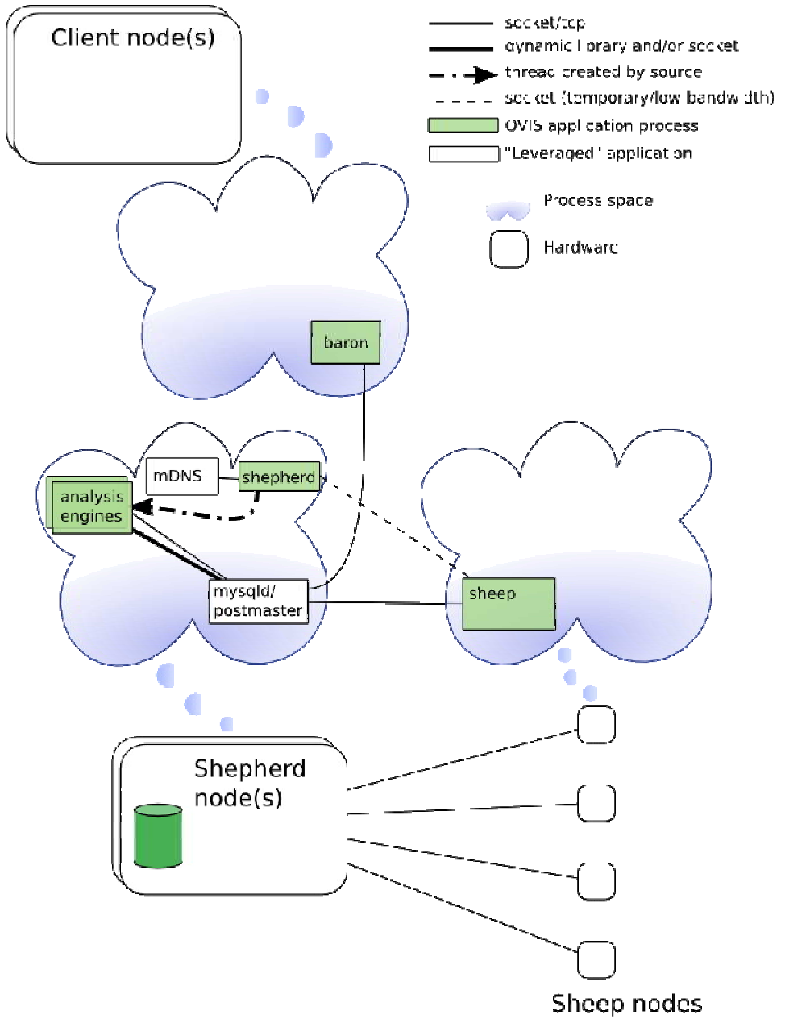
- Impediments
  - Failures are "Rare Events" in small aggregations of machines
  - Data collection mechanisms can cause OS noise which is detrimental in large systems
    - In band (/proc, LM sensors, etc.)
  - Don't know what parameters matter so currently collect as many as possible on as fine a time granularity as possible
    - Data collection scaling issues
      - Storage/retrieval bandwidth
      - Need to store all data for long enough to allow for failure data to be gathered also

Data Acquisition Cont.

- Out of Band (OOB) collection should have minimal impact
  - Separate processor for acquisition and export
  - Separate network for communication
  - IPMI and SNMP can facilitate out of band collection of a subset of desired parameters
- Many parameters are not available OOB
  - /proc -- CPU, Memory, Network, etc.

Enabling Architecture for Data Collection and Fusion

- OVIS is a suite of 3 applications – baron, shepherd, sheep – sharing a common distributed database
- **Baron**: A VTK/Qt user interface (ParaView version in development)
  - Data characteristic exploration
- **Shepherd**: service-node program:
  - Advertises DB availability
  - Responds to requests for analyses
- **Sheep**: a service-node or compute-node program:
  - Listens for shepherds
  - Stores measurements to database on shepherd node



Storage Overview

- OVIS avoids all storage except the database
- Static metadata is stored in an XML file used to initialize a DB
- For scalability, parallel distributed databases are required
  - Analysis requests/results are propagated to all DBs asynchronously
- Some metadata tables are shared between databases
- Parallel databases also imply a parallel baron
  (in development)

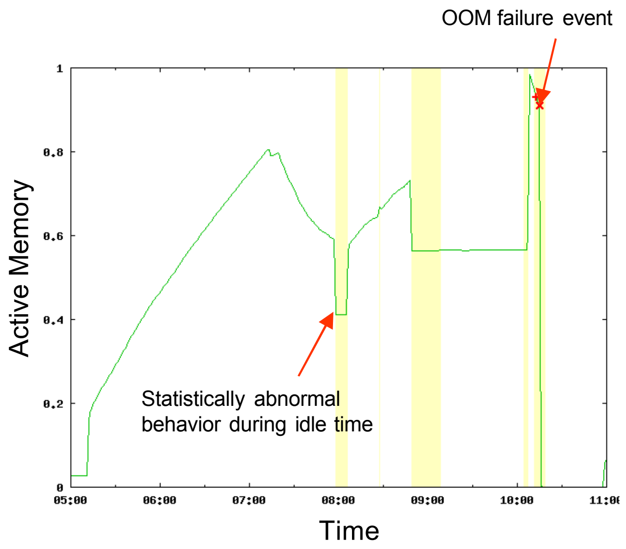
Case Studies

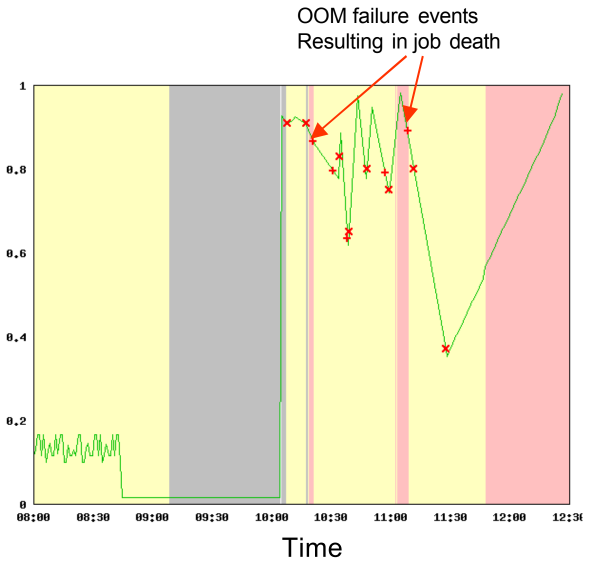
- Red Storm
  - OOB RAS system caused system instability even at a once-per-hour data collection frequency
    - Too course grained
- TLCC
  - Currently collecting on 2SU (288 node, 4K core) system at Sandia Albuquerque (Glory)
    - Hardware related data – once-per-minute
    - Slurm – as it occurs
    - Log file data – before we need it for analysis

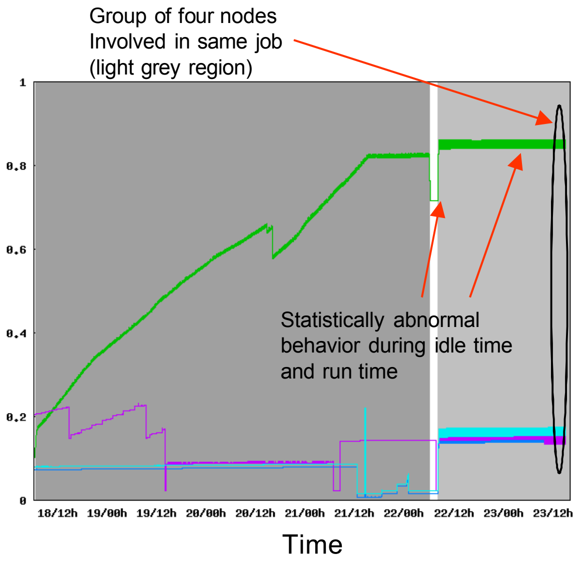
Glory Failures

- We set out to discover precursors of the top three causes of TLCC failures
  - Power Supply
    - Collect various regulated voltages
      - Problems occur upstream of the regulators (caps, inductors, resistors)
      - Need to work with PS designers to understand what it makes sense to instrument
    - Use only a single PS which then runs closer to spec limits
  - Out of Memory
    - Precursor symptoms discovered but root cause yet unknown
  - Stuck CPU
    - Still under investigation

Out of Memory (OOM)







Abnormal Active Memory utilization, detectable during any idle period (yellow shaded region), is seen to persist for hours finally resulting in process being killed due to OOM condition.

Abnormal Active Memory utilization, detectable during any idle period, is seen to result in job failures (pink shaded regions) due to the OOM condition. There are also processes that are killed due to the OOM condition even during idle periods (shaded yellow).

Abnormal Active Memory utilization was detectable both during the idle period (white region) and during the job (right-most grey shaded region). Due to this application's memory requirements the OOM threshold was not reached and this did not result in job death in this case.

Note that Active Memory is represented as a fraction of available system memory (32GB in this case).

Conclusions

- While “black box” approach works for current sized systems it is at best a stop-gap approach
- We are beginning to “pierce the darkness” but still have a long way to go
  - Statistical methods show promise
  - Scalable data collection mechanisms necessary
  - Need more advanced mixed data type analysis methodologies

Future Work

- Cleansed data sets for other researchers to use
- Deployment on additional and larger platforms
- More advanced analyses
- Tighter integration of disparate data sources
- More UI usability features with respect to plotting and graphing

Questions?

https://ovis.ca.sandia.gov

Motivation

- Large amount of time spent/(wasted?) in checkpointing
- Black Box approach to failures in platforms implies scaling wall
  - Best we can do is adjust checkpointing freq based on the measured system average MTTI
- Assumption -- If we understood failure mechanisms
  - Instrument to get precursor warnings where possible
  - Instrument to detect fine granularity failures that could still allow state recovery and hence restart on other resources
  - System and Application interfaces that would facilitate such recovery
  - Know where to invest more up front for higher level of fault tolerance
    - Power supply redundancy cost/fault tolerance tradeoff
    - Memory socket cost/MTTF tradeoff