Nonparametric Multivariate Anomaly Analysis in Support of HPC Resilience

G. Ostrouchov, T. Naughton, C. Engelmann, G. Vallée, and S. L. Scott

Computer Science and Mathematics Division Oak Ridge National Laboratory

Workshop on Computational e-Science at the 5th IEEE International Conference on e-Science

December 7-9, 2009 Oxford, UK This work was partially funded by the Department of Defense and the Extreme Scale Systems Center at ORNL



George Ostrouchov - ostrouchovg@ornl.gov

Collaborators



Stephen W. Poole

Chief Scientist, Computer Science and Mathematics Division LCF System Architect, National Center for Computational Sciences Technical Director, Extreme Scale Systems Center Oak Ridge National Laboratory

Chokchai (Box) Leangsuksun

Associate Professor of Computer Science The SWEPCO Endowed Professor Center for Entrepreneurship and Information Technology Louisiana Tech University





U.S. DEPARTMENT OF ENERGY

LOUISIANA TECH UNIVERSITY®



Equipment for Computational Science: Component Count is Increasing

top500.org processor count:

- About three years ago the entire 500 list broke the million processor mark
- Now the top 7 add up to over a million



World's Most Powerful Computer.

#3

Oak Ridge National Laboratory

"World's Most Powerful Academic Computer"

University of Tennessee

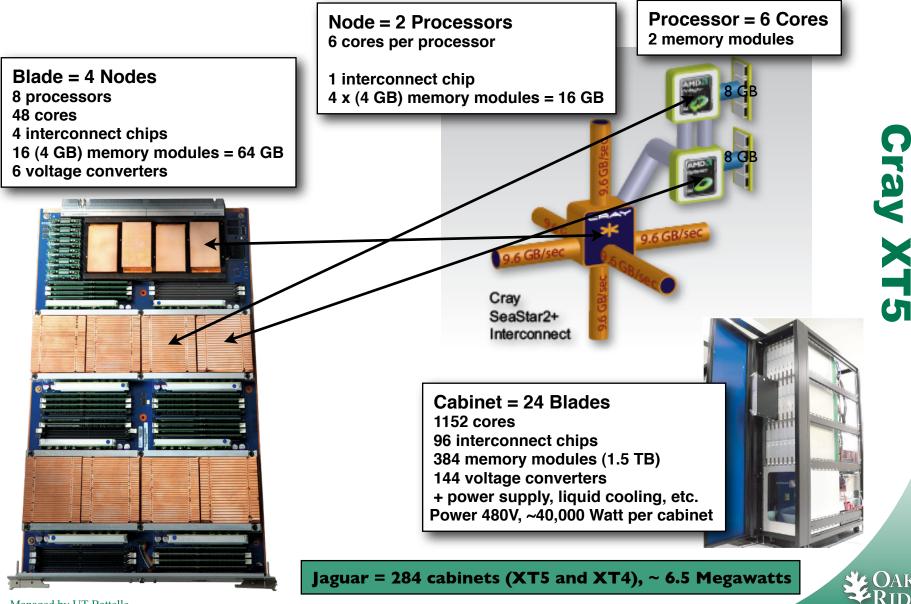
Managed by UT-Battelle for the Department of Energy

Kraken

George Ostrouchov - ostrouchovg@ornl.gov

#1

Jaguar Has Over Quarter Million Cores



Managed by UT-Battelle for the Department of Energy

ational Laboratory

George Ostrouchov - ostrouchovg@ornl.gov

Reactive and Proactive Fault Tolerance

- Reactive: Keeps applications alive through recovery from experienced failures
 - Checkpoint/restart
 - Message logging/replay
 - Effective until failures get too frequent
 - Timely failure reporting for restart
 - Root cause analysis for repair
- Proactive: Keeps applications alive by avoiding failures
 - Preemptive migration
 - Effectively decreases failure rate
 - Needs failure prediction



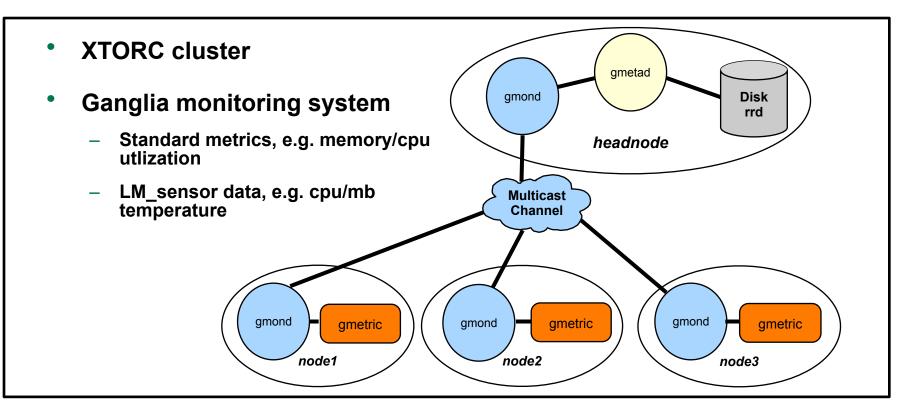
Reactive and Proactive Fault Tolerance

- Reactive: Keeps applications alive through recovery from experienced failures
 - Checkpoint/restart
 - Message logging/replay
 - Effective until failures get too frequent
 - Timely failure reporting for restart
 - Root cause analysis for repair
- Proactive: Keeps applications of the by avoiding failures
 - Preemptive migration
 - Effectively decreases failure rate
 - Needs failure prediction



2

Monitoring Prototype / Data Collection



- Leveraged RRD reader utilities from Ovis v1.1.1
 - RRD readers read tge RRD databases
 - RRD starter start instances of RRD readers on all nodes
 - RRD killer kills all RRD readers
- Format metrics as single line per node



XTORC

Hardware

- Compute nodes: ~45-60 P4 (2 Ghz)
- Head node: 1 P4 (1.7Ghz)
- Service/log server: 1 P4 (1.8Ghz)
- 100 Mb Ethernet

Software configuration

- Spans RedHat 9, Fedora Core 4 & 5
 - RH9: node53
 - FC4: node4, 58, 59, 60
 - FC5: node1-3, 5-52, 61
- RH9 is Linux 2.4
- FC4/5 is Linux 2.6
- NFS mounted /home



Anomaly Detection: Nonparametric Multivariate Analyzer

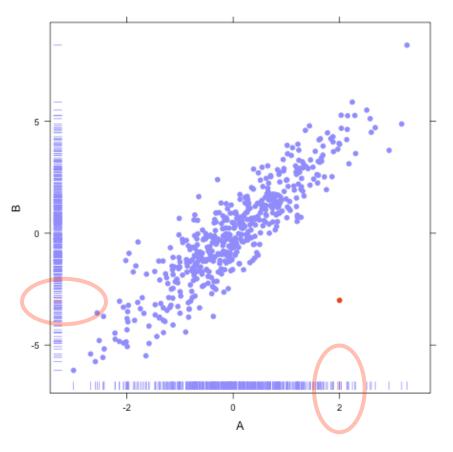
- Ability to view groups of components as statistical distributions
 - Identify anomalous components
 - Identify anomalous time periods
- Based on numeric data with no expert knowledge for grouping
 - Scalable approach, only statistical properties of simple summaries
 - Power from examination of high-dimensional relationships
 - Visualization utility used to explore data
- Implementation uses
 - *R* project for statistical computing



- GGobi visualization tool for high-dimensional data exploration
- With good failure data, could be used for failure prediction



Why do Multivariate Methods Matter?

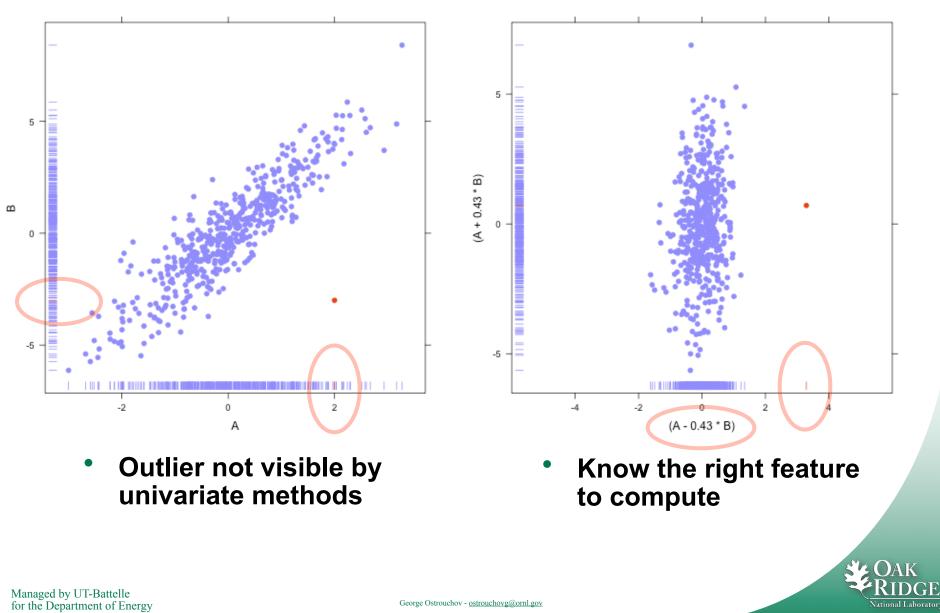


 Outlier not visible by univariate methods

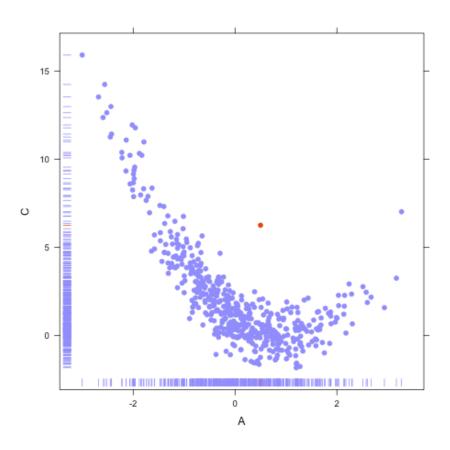


Managed by UT-Battelle for the Department of Energy

Why do Multivariate Methods Matter?

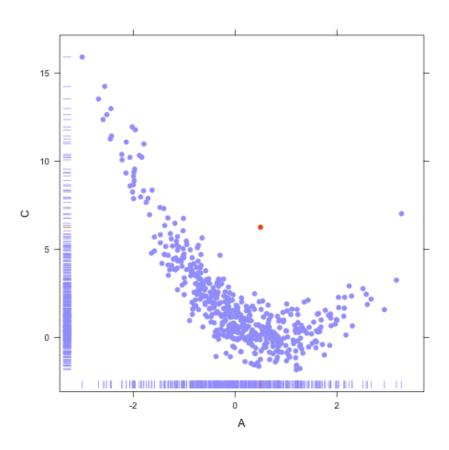


9



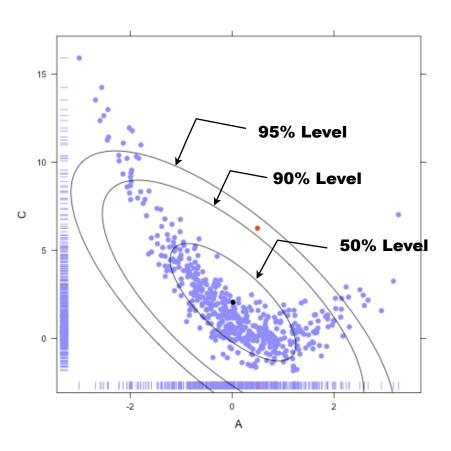




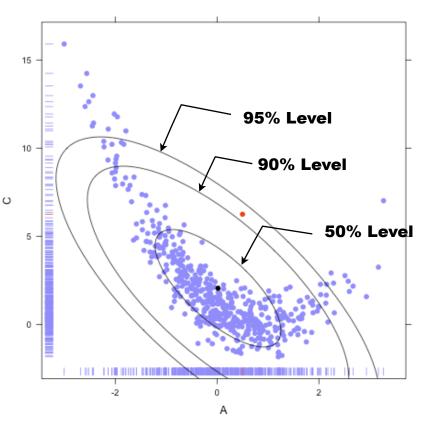




- No linear transformation separates anomaly
- Mahalanobis distance
 - Based on global Gaussian assumption
 - Extremes can have greater distance than anomalies



- No linear transformation separates anomaly
- Mahalanobis distance
 - Based on global Gaussian assumption
 - Extremes can have greater distance than anomalies
- Nearest neighbor distance
 - Related to nonparametric density estimation (Hartigan, 1975)
 - Single Linkage Clustering
 - Minimum Spanning Tree



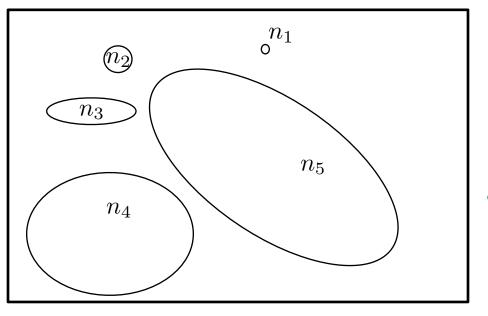
Nonparametric Multivariate Analyzer



Cluster Size Anomaly Measure

- k clusters, sizes n_1, n_2, \ldots, n_k
- Anomaly measure inversely related to cluster size

$$\alpha_i = \frac{1}{n_i}$$



- Sampling from a multivariate density
- Density proportional to probability integral over cluster region
- Measure of cluster anomaly
- Sum over all equal or smaller clusters

$$p_i = n^{-1} \sum_{n_j \le n_i} n_j$$

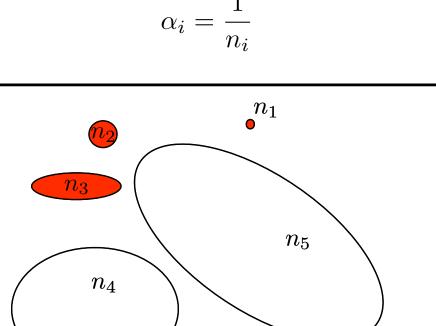
$$\alpha_i = 1 - p_i$$

Interesting to explore as we go down the dendrogram to more clusters



Cluster Size Anomaly Measure

- k clusters, sizes n_1, n_2, \ldots, n_k
- Anomaly measure inversely related to cluster size



- Sampling from a multivariate density
- Density proportional to probability integral over cluster region
- Measure of cluster anomaly
- Sum over all equal or smaller clusters

$$p_i = n^{-1} \sum_{n_j \le n_i} n_j$$

$$\alpha_i = 1 - p_i$$

Interesting to explore as we go down the dendrogram to more clusters



Managed by UT-Battelle for the Department of Energy

Applied to an Old Idle Cluster

untu su Traillica.							
	1	4	-	6		11	
		12h00	5	00h46	10	05h00	
	00h13	1.1	53h00	00n46	01h00		
	00h37		53h01		01h01	05h01	
			53h02	7	01h02	05h02	
	2	12h47	53h03	00h47	01h03	05h03	
	13h47		53h04		01h04	05h04	
		31h00	53h05		01h05	05h05	
			53h06	8	01h06	05h06	
	3		53h07	00h00	01h07	05h07	
	13h00		53h08	00h01	01h08	05h08	
	13h01	31h47	53h09	00h02	01h09	05h09	
	13h02	511147	53h10	00h03	01h10	05h10	
	13h03	39h00	53h11	00h04	01h11	05h11	
	13h04	57100	53h12	00h05	01h12	05h12	
	13h05		53h13	00h06	01h13	05h13	
	13h06		53h14	00h07	01h14	05h14	
	13h07	39h47	53h15	00h08	01h15	05h15	
873 873 474	13h08	57147	53h16	00h09	01h16	05h16	
10	13h09	43h00	53h17	00h10	01h17	05h17	
	13h10	451100	53h18	00h11	01h18	05h18	
	13h11	1.00	53h19	00h12	01h19	05h19	
12	13h12	1.1	53h20	00h12	01h20	05h20	
	13h12		53h20	00h15	01h20	05h21	
200 200	13h14	43h47	53h21	00h16	01h21	05h22	
	13h14	(2) 00	53h23	00h17	01h22	05h23	
	13h16	63h00	53h25	00h18	01h25	05h24	
	13h17	1.1	53h25	00h19	01h24 01h25	05h25	
	13h17	1.1	53h26	00h20	01h25 01h26	05h26	
100	13h19	· · · · · · ·	53h20	00h20	01h20	05h27	
	13h20	63h47	53h27			05h28	
	13h20 13h21		53h28 53h29	00h22	01h28	05h20	
				00h23	01h29	05h20	
	13h22		53h30	00h24	01h30	05h31	
	13h23		53h31	00h25	01h31	05h32	
58 22	13h24		53h32	00h26	01h32	05h33	
	13h25		53h33	00h27	01h33	05h34	
	13h26		53h34	00h28	01h34	05h35	
	13h27		53h35	00h29	01h35	05h36	
	13h28		53h36	00h30	01h36	05h37	
2011 2014 2019	13h29		53h37	00h31	01h37	05h38	
	13h30		53h38	00h32	01h38	05h39	
	13h31		53h39	00h33	01h39	05h40	
	13h32		53h40	00h34	01h40	05h40 05h41	
220	13h33		53h41	00h35	01h41		
2.2 2 2 2 2 2 2 2 2 2	13h34		53h42	00h36	01h42	05h42 05h43	
100 100 100 100 100	13h35		53h43	00h38	01h43		
1075 1075	13h36		53h44	00h39	01h44	05h44	
	13h37		53h45	00h40	01h45	05h45	
204 105	13h38			00h41			
	13h39			00h42			
	13h40			00h43		12	
	13h41			00h44		the remainin	ng
	13h42			00h45		2015 node-	-
	13h43					hours	
	13h44						
	13h45			9			
	13h46			30h07			
2.2.2.2.2.2.2.2.2.2.2.2.2.2.2.2.2.2.2.							

Figure 3 A cluster diagram on left shows the most interesting 400 (out of 2,496) nodehours as rows. Cluster lists are expanded on the right with matching colors.

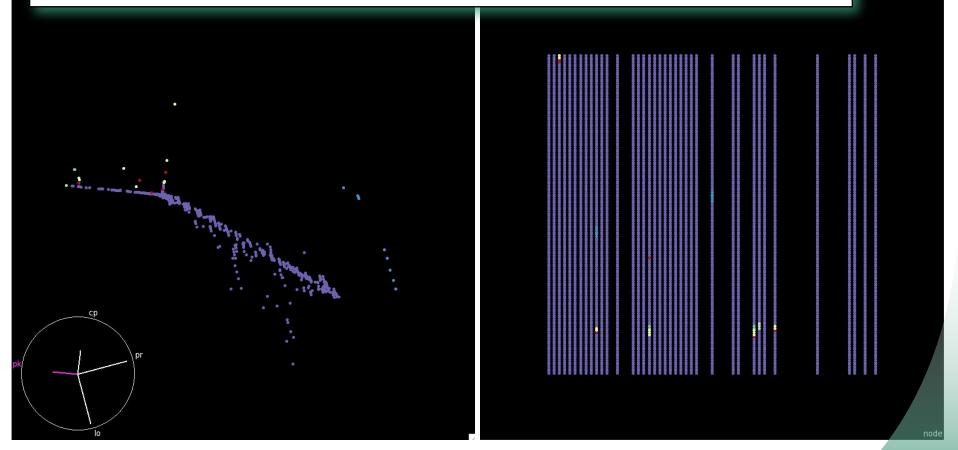
- Node 0 is the most different from the rest, particularly its hours 13, 37, 46, and 47. This is the head node where most services are running. Note that hours 13 and 37 are 24 hours apart, indicating a daily service.
- Node 53 (all hours) runs the older Red Hat 9 while the others run Fedora Core 4/5.
- Nodes 12, 31, 39, 43, and 63 were all down.
- Node 13 was distinct, particularly hour 47, for undetermined reasons.
- Node 1, Node 5, and Node 30 (hour 7) were also distinct for undetermined reasons.



Managed by UT-Battelle for the Department of Energy

Injecting Faults on XTORC

- Clustering node-minute data (R/GGobi)
- Anomaly measure (purple-blue-<u>yellow</u>-red) proportional to cluster size
- Found all injected faults



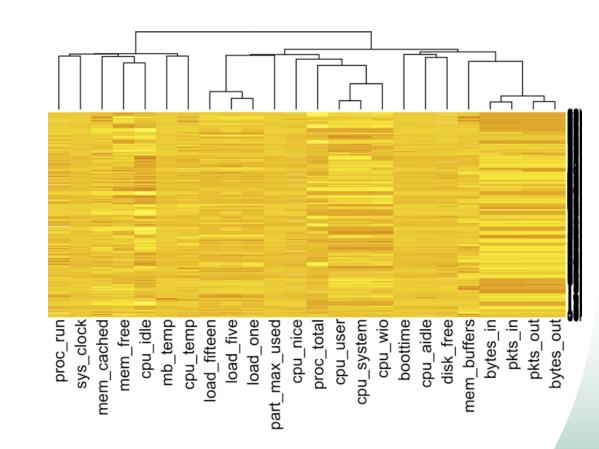
Identifying anomalies: first step to identifying failures and building a failure prediction capability

Managed by UT-Battelle for the Department of Energy

George Ostrouchov - ostrouchovg@ornl.gov

Attributes Can be Clustered for Similarity

- Three groups
 - temperature and memory
 - processor related
 - I/O related
- Finer groups
 - bytes/packets in
 - bytes/packets out
- Gives ordering of attributes



Automation!



Anomaly Detection - Next Steps

Data

- Reduce overhead in data gathering
- Apply to data from Jaguar
- Monitor more fields (ex. HDD)
- Investigate methods to aid data interpretation
- Identify significant fields for given workloads
- Heterogeneous nodes
- Different workloads
 - Base (no/low work)
 - Loaded (benchmark/app work)
 - Loaded + Fault Injection
- Work toward links between anomalies and failures





