### Toward Lightweight, Actionable Analytical Tools Based on Statistical Learning for Efficient System Operations

### **Devesh Tiwari**

Oak Ridge National Laboratory

tiwari@ornl.gov

Thanks to Saurabh Gupta, Christian Engelmann, Sudharshan Vazhkudai, Jim Rogers, Bin Nie, Evgenia Smirni, Franck Cappello, Rinku Gupta, Sheng Di, Marc Snir, Leo Bautista-Gomez, Swann Perarnau, Nathan Debardeleben, Paolo Rech, Don Maxwell, Changhee Jung, Martin Schulz, Ignacio Laguna, and many more smart folks!

This work used the resources of the Oak Ridge Leadership Computing Facility, located in the National Center for Computational Sciences at the Oak Ridge National Laboratory, which is managed by UT Battelle, LLC for the U.S. Department of Energy, under the contract No. DEAC05-00OR22725.

ORNL is managed by UT-Battelle for the US Department of Energy





Courtesy: Rick Slick CUG 2013 Tutorial Slides

#### Hard to store and manage L1 controller syslog-ng Login node rsyslogd Boot node rsyslogd Network node rsyslogd syslog ↓ /SMW queue rsyslogd 🖌 Lustre (LNET) node rsyslogd The syslog queue is only used when (Service nodes) Log files messages cannot be forwarded to the SMW External log host rsyslogd

Courtesy: Rick Slick CUG 2013 Tutorial Slides

	L1 controller
	Spring Watcher - Local Spring Server         Control
	Show Any Severity - from Wark-F4 - last 100 ; manages Update error 10 ; second Updated if 412/00112025/440 Activities     g / Line 2000 Systep Systep Systep Pariod D9 days ImportSearch Resch (0); Server (0
SMW	0         44.270211 51.220.720 AM         321.520.124 head 4         Ondergy         Apr 12.2011000400 stepsic         VID17-730005 UDP respect discorded from 102.142.23(ATE to initial.           0         44.270211 51.220.720 AM         321.520.124 head 4         Ondergy         Apr 12.2011000400 stepsic         VID17-730005 UDP respect discorded from 102.142.23(ATE to initial.           0         44.270211 51.220.1240 AM         321.520.124 head 4         Ondergy         Apr 12.2011000400 stepsic         VID17-730005 UDP respect discorded from 102.142.23(ATE to initial.           0         44.270211 51.220.1241 AM         321.520.124 head 4         Ondergy         Apr 12.201100400 stepsic         VID17-730005 UDP respect discorded from 102.142.23(ATE to initial.           0         44.270211 51.220.1241 AM         321.560.124 head 4         Ondergy         Apr 12.201100400 stepsic         VID17-730005 UDP respect discorded from 102.142.23(ATE to initial.           0         45.270011 52.250.1221 AM AM         321.560.124 head 4         Ondergy         Apr 12.201100400 stepsic         VID17-730005 UDP respect discorded from 152.142.23(ATE to initial.           0         45.270011 52.250.1221 AM AM         321.560.124 head 4         Ondergy         Apr 12.201100400 stepsic         VID17-73005 LDP respect discorded from 152.142.23(ATE to initial.
rsyslogd	4(1)701113231283 AM 231265124 local Anton April 20110030-0 rt-pix VFD-3-32201. How user added to local disess (Inserts admin Pric 2
	4-12/021121312149 34M         321.061124         Korl4         Orbug         Apr12.021100121         trgs         VET-730005: LOP repart disorded from 152.062.210.0176 insid.           4-12/0211213121342364         MM         321.061144         Korl4         Orbug         Apr12.021100120         trgs         VET-730005: LOP repart disorded from 152.062.210.0176 insid.           108 AM         121.061144         Korl4         Orbug         Apr12.00110012:00 in-gas         VET-730005: LOP repart disorded from 152.062.210.0176 in mod.           108 AM         121.061.34         Korl4         Orbug         Apr12.00110012:00 in-gas         VET-730005: LOP repart disorded from 152.062.210.0176 in mod.           108 AM         121.061.34         Korl4         Orbug         Apr12.00110012:00 in-gas         VET-730005: LOP repart disorded from 152.062.210.0176 in mod.           108 AM         121.061.34         Korl4         Orbug         Apr12.00110012:00 in-gas         VET-730005: LOP repart disorded from 152.062.210.0176 in mod.
	Syslog PC (152 109.1.04) Teacher from 152.101.2.35/17 to Exploration 5
External log h	Watcher websers Watcher websers Watche
rsyslogd	SINPSOFT STATE as a second to that the lag reserver is state or whethered. Excemmended Action 5

# Hard to store and manage

, SMW	2 4,12/2011 1.32.31.729 AM 4,12/2011 1.32.32.349 AM 4,12/2011 1.32.32.349 AM 4,12/2011 1.32.31.814 AM 4,12/2011 1.32.31.814 AM 4,12/2011 1.32.31.811 AM	101.5621.64 head4 101.5621.64 head4 101.5621.64 head4 101.5621.64 head4 101.5621.64 head4	Dobug Dobug Dobug Dobug Naming	Apr 12 2011 (0.08.50 strpin Apr 12 2011 (0.08.61 strpin Apr 12 2011 (0.08.61 strpin Apr 12 2011 (0.08.60 strpin Apr 12 2011 (0.08.60 strpin Apr 12 2011 (0.08.60 strpin	UE: 7 2005 UE9 report disorded from 102162.23(27) to inid. UE: 7 2005 UE9 report disorded from 102162.23(27) to inid. UE: 7 2005 UE9 report disorded from 102162.23(27) to inid. UE: 7 2005 UE9 report disorded from 102162.23(27) to inid. UE: 7 2005 UE9 report disorded from 102162.23(27) to inid. UE: 7 2005 UE9 report disorded from 102162.23(27) to inid. UE: 7 2005 UE9 report disorded from 102162.23(27) to inid.
rsvslogd 🖌	4/12/2011 1:12:21 310 AM	131.168.1.54 local4	Nation	Apr12 2011 00:00:40 st-pix	%FD-5-502001: New user added to local dates: Uname: admin Print 2
	4/12/2013 1:11:44,008 AM	131.168.1.54 local 4	Debug	Apr 12 2011 00:00:00 it -pix	SPD-7-21005: UDP regent disarded from 152.052.231/137 to ined.
	4/12/2013 1:31:13.445 AM	131168.1.54 local4	Debug	Apr12.3011.00:07:31 st-pix	SPD-7-710805: UDP request decarded from 152168.2.31/137 to ined
	4/12/2011 1-11-12-588 AM	191168.1.54 local4	Debug	Apr12.2011.00:07:50 it-pix	SPD-7-710105: LDP request decarded from 182168.2.33/137 to insid.
	LES AM	131.165.1.54 local4	Debug	Apr12 2011 00:07:50 nt-pix	SPD-7-710805: UDP request discarded from 152168.2.31/137 to insid.
	STREET, STREET	197.0663.54 Ional4	Debug	Anr173011001539 (6-09	SPB-7-7000110P report downord from 10/1602/07/10/10 to made.
External log h	Syslog Watcher	(170L108-L04)) west discarded from t box-rail	192 108 2 35/1 Tet 2	37 to Exploration This reasons that covers in mention as a DWD no. Records of Records of Records of Records of the Records of	Trendry, April 12, CELL LOJ 44 605 M parageness when the associativy applicates does not have a LOG approx- tion EDP research. The message issue and have a LOG application to EDP research. The message issue and have a State of the top operating the top applicable, each 3 is a form on another and participation that an employee a substantial top and the top application of the top applicable, each 3 is form on another advance of the message. The message association measurement of a time every 10 top the top applicable, each 3 is form on another advance of the top the top applicable, each 3 is form on another advance of the top the top applicable on the top applicable on the top applicable on the top the top applicable on the top applicable on the top applicable on the top the top applicable on the top applicable on the top applicable on the top the top applicable on the top applicable on the top applicable on the top the top applicable on the
	Syslog Wat		W. S.	vsloawa	tcher

### Hard to store and manage

L1 controller

### Accurate interpretation hard

SMW rsyslogd	(* 127031 5120110 500 (* 127031 5120110 80 AM (* 127031 5120110 80 AM (* 127031 5120110 81 AM (* 127031 512010 81 AM (* 127031 5120	10136134 ben14 10136134 ben14 10136134 ben14 10136134 ben14 10136134 ben14 10136134 ben14 10136134 ben14 10136134 ben14 10136134 ben14	Dahay Dahay Dahay Dahay Nation Nation Nation	Apri2 2011 000000 wijes Apri2 2011 000041 wijes Apri2 2011 000040 wijes	401.7.7005.01 901.7.70065.01 901.7.70065.01 901.7.70065.01 901.4.40100.14 901.4.502201.00 901.4.502201.00 901.4.502001.00 901.4.502001.00	20° response discondust forces 1822 DP response discondust forces 1822 Related communities discondust for Related communities discondust for Related communities discondust forces 1822	48223(43716 mml), 68223(43716 mml), 68223(43716 mml), 68223(43716 mml), duringed state to up name admin Pric 2, nn PDi comole 68223(43716 mml),
Loc		201200124 Nor44 201200124 Nor44 201200124 Nor44 201200124 Nor44 201202100124	Debug Debug Debug Debug	Apr12.2011.003730 tt-pix Apr12.2011.003730 tt-pix Apr12.2011.003730 tt-pix Apr12.2011.003730 tt-pix	SPD-7-710105-10 SPD-7-710105-10 SPD-7-710105-10 SPD-7-710105-10	OP request disorded from 1821 OP request disorded from 1821 OP request disorded from 1821 OP request disorded from 1821 Towaday, April 12, 201	68,233,137 to Hed. 68,233,137 to Hed. 68,233,137 to Hed. 68,233,137 to Hed. 11 L (31,44,000 AM
External log h	Watcher	spuest discorded from 15	2.168-2.35/1	37 to Explanation This reason that service that device in metation is an DVMP re- When the s records so	b) ge appears when the s the SEP request, to belong to any se- opears light the ser- part with an array, even is easy, the that the log receive that the log receive	a neutrity appliance does not the remaining can ame indicate action on the security appliance vice same) when the security papies, even if it is from a remaining excession a measure of remaining excession a measure of remaining excession and more relations.	have a LOP sever a TCP packet a. In addition, this spoken review authorized heat, of a time every 10
rsyslogd				Recommen	OHE ACTION >		







### System failures exhibit temporal and spatial locality.

![](_page_8_Figure_1.jpeg)

## Observation holds true across systems and failure types, consistently across long range of periods.

### System failures exhibit temporal and spatial locality.

![](_page_9_Figure_1.jpeg)

Observation holds true across systems and failure types, consistently across long range of periods.

S

S

S

![](_page_10_Figure_1.jpeg)

S

![](_page_11_Figure_1.jpeg)

![](_page_11_Figure_2.jpeg)

![](_page_12_Figure_1.jpeg)

#### Refer to the paper for model validation and simulation results

[DSN 2014] Lazy Checkpointing: Exploiting Temporal Locality in Failures to Mitigate Checkpointing Overheads on Extreme-Scale Systems Devesh Tiwari, S Gupta, S Vazhkudai, IEEE/IFIP Int'l Conference on Dependable Systems and Networks (DSN), 2014.

![](_page_13_Figure_1.jpeg)

[DSN 2014] Lazy Checkpointing: Exploiting Temporal Locality in Failures to Mitigate Checkpointing Overheads on Extreme-Scale Systems Devesh Tiwari, S Gupta, S Vazhkudai, IEEE/IFIP Int'l Conference on Dependable Systems and Networks (DSN), 2014.

### Idea: On job restart or a new job allocation a fraction of compute capacity is not utilized (quarantined)

#### **Quarantine Granularity**

Quarantine		
ann ann ann ann Dona Dona Dona ann ann ann ann ann ann ann ann ann	and and and and and and and and and and	Jop Jop Jop Jop
		Job

Fraction of avoided system failures versus compute resource waste

#### **Quarantine Time Duration**

![](_page_14_Figure_6.jpeg)

Diminishing returns on the number of avoided failures

#### System Utilization vs. Reliability

![](_page_14_Figure_9.jpeg)

Trading-off lower system utilization for improved reliability

[DSN 2015] Understanding and Exploiting Spatial Properties of System Failures on Extreme-Scale HPC Systems Saurabh Gupta, Devesh Tiwari, Chris Jantzi, Jim Rogers, Don Maxwell, IEEE/IFIP Int'l Conf on Dependable Systems and Networks (DSN), 2015.

![](_page_15_Figure_1.jpeg)

![](_page_16_Figure_1.jpeg)

■ % Failures Avoided ■ % Quarantine Hours

![](_page_17_Figure_1.jpeg)

■ % Failures Avoided ■ % Quarantine Hours

![](_page_17_Figure_3.jpeg)

![](_page_17_Figure_4.jpeg)

Significant fraction of failures can be avoided from interrupting production applications

Debug or non-production jobs can be scheduled on quarantine nodes

### Feedback Loop for ModSim Community

![](_page_18_Figure_1.jpeg)

These insights can potentially change the way we design fault-injection modsim frameworks, operate production machines, and plan for future systems.

### **Opportunity for Actionable Analytical Tools**

![](_page_19_Figure_1.jpeg)

Workload and system generated data

**Environment and facilities data** 

How can we (machine learning) fuse all this data to take meaningful, timely, and profitable decisions at-scale?

### **My Personal View**

Future large scale system will have heterogeneity in terms reliability levels, too.

Parts of large systems will go in transient lower reliability, degraded performance, and large performance variability modes.

# Traditional "replace and continue" approach will not be sustainable.

We will need theoretically-sound techniques and tools to "dynamically" <u>manage</u> this new kind of heterogeneity.

# Traditional "replace and continue" approach will not be sustainable.

We will need theoretically-sound techniques and tools to "dynamically" <u>manage</u> this new kind of heterogeneity.

Denial and blame shifting will continue to work for some time in near future. ③

### **Thanks!**

### **Devesh Tiwari**

Oak Ridge National Laboratory

tiwari@ornl.gov

devesh.dtiwari@gmail.com