

Scaling Performance Analysis Tools to Hundreds of Lustre Filesystems

Ellis Wilson - Microsoft

Today's Talk

- \cdot The Need for Advanced Logging, Metrics, and Alerting
 - Problem statement for Azure Managed Lustre Filesystem (AMLFS)
- \cdot Making Sense of Logs from Thousands of Nodes
 - $\cdot\,$ Collection and Aggregation of node syslog and other log output
 - · Azure Monitor Log Curation, Search, and Analytics
- \cdot Node and Cluster Performance Metrics
 - $\cdot \,$ What we collect, how, and why
 - · Overview of the Azure Monitor Metrics Interface(s)
- \cdot Health Monitoring and Alerting
 - \cdot Automated problem detection (before the customer notices)
 - Azure Monitor Health Interface

The Need for Observability in Lustre

- Typical end-user feedback tends to be ... sparse:
 - "It's not working"
 - "It's stuck"
 - "It's slow"
 - · "I get an error"
- Lustre stats are useful, but only available at the CLI on-box

- A lot more to monitor:
 - Networking health and performance
 - \cdot CPU load

...

- Memory utilization
- Disk throughput, latency, load, capacities
- · Client connectivity
- · Cluster-wide health (e.g., heartbeats)
- System infrastructure logs
- Userspace crashes and core files
- \cdot Kernel panics and vmcores
- Lustre-related application health
- Various versions (distro, Lustre, software)

Now Scale that to Thousands of Clusters

- Azure Managed Lustre (AMLFS) needed to design logging, metrics, and health telemetry to support tens of thousands of Lustre nodes
 - $\cdot \,$ One admin per cluster is not even remotely viable
- · Getting on-box to support customers is a non-starter
 - $\cdot\,$ Ignoring scaling issues, privacy and security requirements preclude this
 - · Cluster may be gone transient/job-based usage
 - \cdot Must export (solely non-sensitive) logs, metrics, and health information to centralized service
- \cdot Need support for powerful querying
 - $\cdot\,$ Downloading logs from ten thousand nodes and grepping won't cut it
- · Visualize performance and health metrics
 - $\cdot\,$ No sane way to find the one problematic OSS in a cluster without this
- · Internal and Customer-facing Alerting

The Azure Monitor Data Platform



https://learn.microsoft.com/en-us/azure/azure-monitor/overview

AMLFS Logging Design Criteria

- Ingest system logs from every AMLFS-side cluster node
 - MGS, MDSes, OSSes, and HSM Agents
 - All syslog traffic
- New log messages uploaded and available within low-digit minutes
- No personally-identifiable information (PII) exfiltrated
 - E.G.: File names or data
- Support powerful querying available in Azure Monitor
 - Syslog messages quantized into
 - Time, PID, ProgramName, Facility, Severity
 - Additional metadata associated with each log
 - Region, Cluster ID, Hostname, Role, RoleInstance
- Enable non-syslog log ingest for metric-like data

Azure Monitor Log Querying Interface: Dgrep

- \cdot Dgrep supports server and client-side queries
 - Typical debugging pattern:
 - \cdot Server-side selection of a cluster or a nodes logs over a time period
 - $\cdot\,$ Client-side filtration down to the program, PID, or message content of interest
- · Much more powerful queries become possible:
 - Example: Search all OSSes in the fleet for a specific Lustre error message over the last 3 days
 - $\cdot\,$ Example: Gather logs from copytools on primary agents in a specific region
 - \cdot Example: Gather logs for a specific agent over the last 30 days at or above warning severity
- · Deeper log analysis becomes possible with aggregates
 - \cdot Average, Count, Max, Min, and Sum across all previously mentioned dimensions
 - Example: Count by RoleInstance in a large cluster to find a misbehaving node
 - $\cdot\,$ Example: Count by Msg to find log spam
 - Example: Sum Severity by Time and sort to find times with lots of warnings/errors

Dgrep User Interface: Simple Search by Time/Cluster

	Account >				Docs 🖸	▲ 🕤 (9?	(
	🔍 Server Query 🝃 😫 🌞 🗞	▼ Client Que	ery			м	QL KQL	Q
		1 sour	rce					
		2 so	ort by PreciseTi	imeStamp asc				
_								
	Show Azure security pack events							
-	Time range							
~~	Mow 05/01/2024 16:12 UTC							
_	-5mins -1 min +1 min +5mins							
	± + - 1 Hours -	I Logs	€ ± Fine	d prefix with '-' to negate	General Chart	II A	ggregates	
6	1 2 3 5 15 30	Role	RoleInstance	EventTime	Msg	ProgramName	PID	
	Scoping conditions	mdsmgs	mdsmgs0000	05-01-2024 15:13:03	AzureQueueMonitor: Sent message '{'payload': {'clusterstatus': '	laaso-azure	1575	Â
=	Add Index	mdsmgs	mdsmgs0000	05-01-2024 15:13:13	2024/05/01 15:13:13.217978 INFO: [heartbeat] Start	azsecd	1574	
v	Select a comparand, operator, and value or values for each condition you require.	mdsmgs	mdsmgs0000	05-01-2024 15:13:13	2024/05/01 15:13:13.226116 INFO: Read config file	azsecd	1232296	
	Tenant • == • ×	mdsmgs	mdsmgs0000	05-01-2024 15:13:13	2024/05/01 15:13:13.226719 INFO: Added pid '1232296' to cgroup '	azsecd	1232296	
	★ cde226a3-dd6c-44f1-91c3-2e08b24c0699	mdsmgs	mdsmgs0000	05-01-2024 15:13:13	2024/05/01 15:13:13.226916 INFO: Starting 'heartbeat' scan	azsecd	1232296	
-		mdsmgs	mdsmgs0000	05-01-2024 15:13:14	2024/05/01 15:13:14.457549 INFO: Scan 'heartbeat' completed	azsecd	1232296	
		mdsmgs	mdsmgs0000	05-01-2024 15:13:20	{'qmonrx': 1, 'qmontx': 15541, 'qmonerr': 0, 'busrx': 15538, 'bu	laaso-azure	1575	
?	New condition for Field	oss	oss0000	05-01-2024 15:58:51	Sending event source=amlfs_node_info	laaso-AmlfsN	2820	
	Filtering conditions Simple MOL KOL	oss	oss0000	05-01-2024 16:04:18	{'qmonrx': 1, 'qmontx': 5, 'qmonerr': 0, 'busrx': 2, 'buserr': 0}	laaso-azure	2264	
	Simple Mee Ree	oss	oss0000	05-01-2024 16:07:13	Configuration (uniq=d0e799bd-2947-488a-bbe7-3dc7617fafad) has no	laaso-config	2265	
-	Select a comparand, operator, and value or values for	mdsmgs	mdsmgs0000	05-01-2024 16:08:06	AzureQueueMonitor: Sent message '{'payload': {'clusterstatus': '	laaso-azure	1575	
-		mdsmgs	mdsmgs0000	05-01-2024 16:08:24	{'qmonrx': 1, 'qmontx': 15596, 'qmonerr': 0, 'busrx': 15593, 'bu	laaso-azure	1575	
	Tieu	oss	oss0000	05-01-2024 16:08:52	Sending event source=amlfs_node_info	laaso-AmlfsN	2820	
4	The soft-launch of Schrems II	mdsmgs	mdsmgs0000	05-01-2024 16:09:06	AzureQueueMonitor: Sent message '{'payload': {'clusterstatus': '	laaso-azure	1575	
	enforcement is in July 2023.	oss	oss0000	05-01-2024 16:09:19	{'qmonrx': 1, 'qmontx': 5, 'qmonerr': 0, 'busrx': 2, 'buserr': 0}	laaso-azure	2264	
	Click here to learn more	mdsmgs	mdsmgs0000	05-01-2024 16:10:06	AzureQueueMonitor: Sent message '{'payload': {'clusterstatus': '	laaso-azure	1575	
	Miscellaneous 🗸	mdsmøs ∢	mdsmps0000	05-01-2024 16:10:51	laaso.filesvstem.ostning[3035]: ost ning daemon is still alive			•

Metrics via Logs

- Client export stats are also collected as logs as they don't fit well into our normal metrics infrastructure
 - Great for locating problem clients or doing deep analysis on why "It's running slowly today"

PreciseTimeStamp	Role	RoleInstance	target	op_name	samples	min	max	sum
05-01-2024 16:52:03	oss	oss0001	obdfilter.lustrefs-OST0001.exports.10.17.0.8@tcp.stats	statfs	11	1	71	133
05-01-2024 16:52:03	oss	oss0001	obdfilter.lustrefs-OST0001.exports.10.17.16.100@tcp.stats	no-op-data	0	0	0	0
05-01-2024 16:52:03	oss	oss0001	obdfilter.lustrefs-OST0001.exports.10.17.16.101@tcp.stats	no-op-data	0	0	0	0
05-01-2024 16:52:03	oss	oss0001	obdfilter.lustrefs-OST0001.exports.10.17.16.102@tcp.stats	no-op-data	0	0	0	0
05-01-2024 16:52:03	oss	oss0001	obdfilter.lustrefs-OST0001.exports.10.17.32.5@tcp.stats	quotactl	60	1	4037	332
05-01-2024 16:52:03	oss	oss0001	obdfilter.lustrefs-OST0001.exports.10.17.32.6@tcp.stats	write_byte s	2	1	4194304	6667
05-01-2024 16:52:03	oss	oss0001	obdfilter.lustrefs-OST0001.exports.10.17.32.6@tcp.stats	write	2	4	652866	7114
05-01-2024 16:52:03	oss	oss0001	obdfilter.lustrefs-OST0001.exports.10.17.32.6@tcp.stats	set_info	5	0	57	39
05-01-2024 16:52:03	oss	oss0001	obdfilter.lustrefs-OST0001.exports.10.17.32.6@tcp.stats	quotactl	54	1	93	324
05-01-2024 16:52:03	oss	oss0001	obdfilter.lustrefs-OST0001.exports.10.17.32.7@tcp.stats	quotactl	12	1	3275	157
05-01-2024 16:52:03	mdsmgs	mdsmgs0000	mdt.lustrefs-MDT0000.exports.10.17.0.4@tcp.stats	no-op-data	0	0	0	0
05-01-2024 16:52:03	mdsmgs	mdsmgs0000	mdt.lustrefs-MDT0000.exports.10.17.0.5@tcp.stats	no-op-data	0	0	0	0
05-01-2024 16:52:03	mdsmgs	mdsmgs0000	mdt.lustrefs-MDT0000.exports.10.17.16.100@tcp.stats	no-op-data	0	0	0	0
05-01-2024 16:52:03	mdsmgs	mdsmgs0000	mdt.lustrefs-MDT0000.exports.10.17.16.101@tcp.stats	no-op-data	0	0	0	0
05-01-2024 16:52:03	mdsmgs	mdsmgs0000	mdt.lustrefs-MDT0000.exports.10.17.16.102@tcp.stats	close	1	3	285576	22

AMLFS Metrics Design

- Metrics are the second pillar of our approach to observability
 - Used for both performance analysis and cluster health triage
- Daemon metrics process on every node in every cluster collects, processes, and sends metrics to Azure Monitor Metrics
- Metrics collected at varying intervals, and via different means
 - Some metrics are gathered by running utilities (e.g., lctl, iostat)
 - Others gathered more directly via Python libraries (e.g., psutil)
- Similar time between upload and visibility to Logs (minutes)
- Two interfaces available for visualizing metrics
 - Jarvis and Grafana

Component Metrics

· CPU

- Overall percentages broken down akin to top (busy, idle, iowait, etc)
- · Per-core percentages

· Memory

 Capacity in various states (total, free, available, cached, slab, etc)

\cdot Networking

- · Total packets in/out
- Error counts

· Disk Performance

- Throughput, utilization, merges, iops, request sizes
- \cdot OS/Crash/Log Disk Capacities
 - Fullness by bytes (total, used, free, %)

\cdot Data Disk Capacities

- Fullness by bytes (total, used, free, %)
- Fullness by inodes (total, used, free, %)

Lustre-specific Metrics

\cdot OSS:

- \cdot Request statistics by request type
 - · Total requests
 - $\cdot \,$ Total bytes moved in requests
 - $\cdot\,$ Min/max op size since restart
 - $\cdot\,$ Min/max op latency since restart
- \cdot Total evicted clients
- Total connected clients

\cdot All Cluster Nodes

- Event Alert (e.g., node restart)
- · AMLFS Heartbeat
- AMLFS Version

\cdot MGS/MDS:

- · Same as OST, plus:
- HSM request count by type (restore, remove, archive)
- HSM current/completed/errored requests
- \cdot HSM registered agents
- · Changelog unread events
- · Changelog size
- LDLM MGS Timeouts

Geneva Dashboard Interface: Cluster Diagnostics

ClusterOverview			🌀 View in Grafana 🕂 Widget 🖻 Share	🖒 Refresh 🔚 🍸 🚳 🕂
Time: 05/01, 05:46	PM to 05/03, ∨ Region ∨ Tenant: 941672dc-32	203-4cd6-a5a5 🏏		🔓 Parameters 🛛 葦 Overrides
i This account has of Jarvis dashbo	s a Grafana workspace at <u>https://amlfs-grafana-devtest-ggawaef4</u> bards.	<u>b2bvfeh5.eus2.grafana.azure.com/</u> . Consider using that for	your dashboarding needs henceforth. <u>Please click here to le</u>	<u>earn more</u> about using Grafana instead
OSS - ost on byte	s total - total cluster READ throughout (bytes/minute)		es total - total cluster WRITE throughput (hytes/minute)	- @ (() اب
2,500G	Thursday, May 02, 2024 21:30 • 941672dc-3203-4cd6-a5a5-e6dc6e103bea :: ;	2.73T Thursday, May 02, 2024 15:30 • : 1.00		
1,000G 500G		1,000G 500G		
0 2.N	May 08:00 16:00 3. May 08:0	0 16:00 4. May 2. 1	May 08:00 16:00 3. May	08:00 16:00 4. May
All None	— 941672dc-3203-4cd6-a5a5-e6dc6e103bea :	All None		3bea :

Geneva Dashboard Interface: Disk Stats for One Node



Geneva Dashboard Interface: HSM Metrics



Grafana Dashboard Interface: Per Node Stats

Ø		Q Se	earch or jump to	≡ ctrl+k	•		0	ም 🔮	-
\blacksquare Home > Dashboards > AMLFS > N	Node Diagno	stic 🔂 📽		· ۵	② 2024-05-02 13:18:00	to 2024-05-02 15:59:57 🗸) Q G	~ /	
- neennaa - neennat - neennat - n	ee007d 📥 oe	:enn54 — acenn42 — acenn2h — a	กระกิกรี่ส		- ass0002 - ass0054 - a	neeNN?f 🛑 neeNN5d 🛑 neeNI	076 - 000079		
MDS ops/sec - cluster-wide				MDS Latency By OpType	e (msec)				
100.00 ops/s	-			150.00					
75.00 ops/s				100.00					
50.00 ops/s	,								
25.00 ops/s				50.00					
0.00 ops/s 13:30 13:45 14:00	14:15 14	30 14:45 15:00 15:15 15:	:30 15:45	0.00					
- close - getattr - getxattr - open crossdir_rename - rmdir - rename	🗕 mknod 🗕	• UNLINK — sync — setattr — stat	tfs 🗕 mkdir	13:18:00 13:45 - statfs - getxattr -	5:00 14:12:00 getattr — open — unlink —	14:39:00 15:06:00 mknod — setattr — sync	15:33:00 — close — mkdir		
Mean op latency - cluster-wide averages	6								
Display side by side latency graphs									
OSS Ops/sec 🛈		OSS Latency By OpType (msec)		MDS Ops/sec 🕕		MDS Latency By OpType	(msec) 🛈		
read 16.66	ôK ops/s	GET_info	0.00 ms	close	12.05 ops/s	mkdir	с).00 ms	
	A			MM_AMAJ	Irhhlanna				
write 13.66	ôK ops/s	sync	22.65 ms	getattr	6.89 ops/s	statts	C).00 ms	
AAA AAA				getxattr	6.55 ops/s	open	C).05 ms	
sync 35.7	78 ops/s	destroy	96.87 ms						
				open	6.02 ops/s	rename	C).00 ms	
destroy 35.7	78 ons/s		3 0 1			aetvattr		00	
	0 005/5		3.04 ms			getAutt	L L	9.00 ms	
	0	write	00.00	UNLINK	5.96 ops/s	getattr	(0.01 ms	
25.0	ops/s		90.23 ms						

Cluster Health Monitoring and Alerts

- Logs and metrics are great when you know you have a problem
 - But this tends to occur after an angry Lustre user has escalated a complaint
 - Which means it's guaranteed to be 3am
- Knowing there is a problem prior to or simultaneously with a customer being made aware is infinitely better
- Health Monitors and Alerts are the third pillar of observability
- Most of this leverages existing metrics discussed before
 - Take existing stats that we're already sending to Azure Monitoring, and set rules on them
 - If a stat is found to fail the test so many times in a row or for known duration, automatically create an incident ticket and autopopulate information about the problem
 - Engineers can jump right from the incident via links to logs and metrics around that time to triage/diagnose

AMLFS Monitors

- We have rules that raise alerts and/or create incidents of varying severities for the following:
 - Missing Heartbeats
 - Heartbeat indicates Degraded cluster
 - Unexpected reboot/shutdown events
 - MGT/MDT/OST capacity available too low
 - \cdot Non-data disk capacity available too low
 - Coredumps observed
 - $\cdot\,$ Read and Write op latencies too long
 - Total CPU percentage too high
 - $\cdot\,$ Free memory too low
 - \cdot We continue to adjust and add more as customer issues find cases we'd like to detect early
- \cdot Most of these only "pop" if they hit many times over some duration

Azure Monitor Health Interface: Cluster Getting Full

AZSCmdmprod	irces 🔻		1hrs 2hrs 6hrs	12hrs 24hrs O	UTC From 🛗 04	/15/2024 04:00	To 🛗 Now 0	5/02/2024 03:11	
							🧔 Try i	n Grafana 🔍) C
(i) This account has a Grafana instance as: using that henceforth. Learn more here	sociated with it at https://# 2.	MLFS-Graf	ana-hpf9a4cwayhaa5fg.	eus2.grafana.azure.com	. This instance comes w	ith an out of the box	x dashboard that mimics t	the Jarvis health pag	e. Consider
04	4:00	04/18 -	12:00	04/21 - 21:00	04/25	- 06:00	04/28 - 15:00		0 03:11
oss_capacity_monitor									••••
 XAMLFS-cluster-oss-filesyst 				(••••
Evaluated every 00:03:00 against the last 00:10:00	ost_filesfree								
	80M								
	-80M		20 4.5		24 4.5		20 4 5		2.14
	ost kbytesavail	18. Apr	20. Apr	22. Apr	24. Apr	26. Apr	28. Apr	30. Apr	2. May
	4. <u>8G</u>								
	-4.8G 16. Apr	18. Apr	20. Apr	22. Apr	24. Apr	26. Apr	28. Apr	30. Apr	2. May
	ost_filesfree_samples								
	10.2								
	7.8	19 Apr	20 Apr	22 Apr	24 Apr	26 Apr	28 Apr	20 Apr	2 1424
	ost_kbytestotal_samples	To. Apr	20. Apr	22. Apr	24. Apr	20. Apr	zo. Apr	SU. Apr	2. May
	10.2								
	7.8								
	16. Apr	18. Apr	20. Apr	22. Apr	24. Apr	26. Apr	28. Apr	30. Apr	2. May

Wrapping Up

- In AMLFS we had to design an observability infrastructure to support thousands of clusters, and hundreds of thousands of cluster nodes, including nodes that may not be around when you need to debug
- Logs, Metrics, and Alerts comprise our three pillars of observability
 - Support Log aggregation, parsing, querying, and statistics via Azure Monitor Logs/Dgrep
 - Support Metric gathering and aggregation from various on-box utilities and visualization
 - Support Monitors/Alerts and automated incident creation when certain metrics go south
- In most cases, we can avoid getting on-box until we know exactly what we plan to do to fix the problem
 - Less time on-box, less room for user-error, far less exposure to customer data
- Again, looking for feedback and others experiences in this area!



Thanks! Questions?