OAK RIDGE National Laboratory

HPCM Monitoring Experience @ OLCF

Ryan Adamson

(and Tim Osborne, Rachel Palumbo, Corwin Lester)

National Center for Computational Sciences (NCCS)

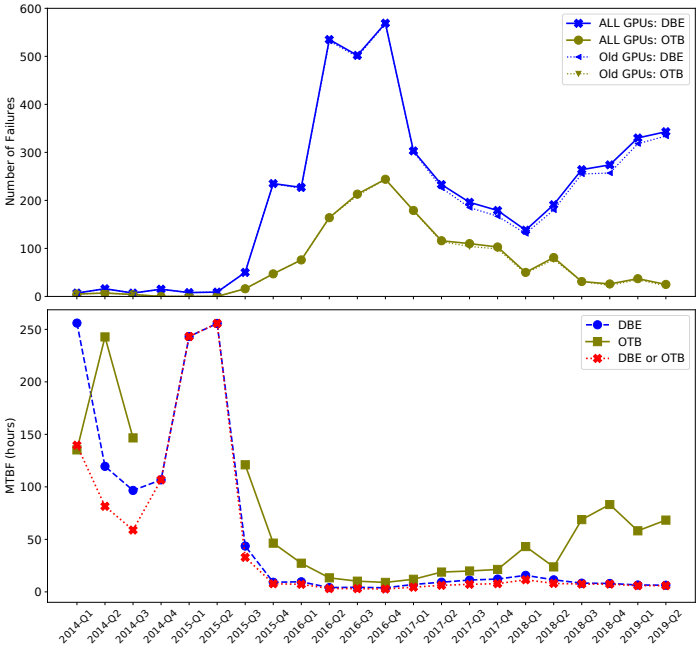
Oak Ridge National Laboratory (ORNL)

CUG 2023

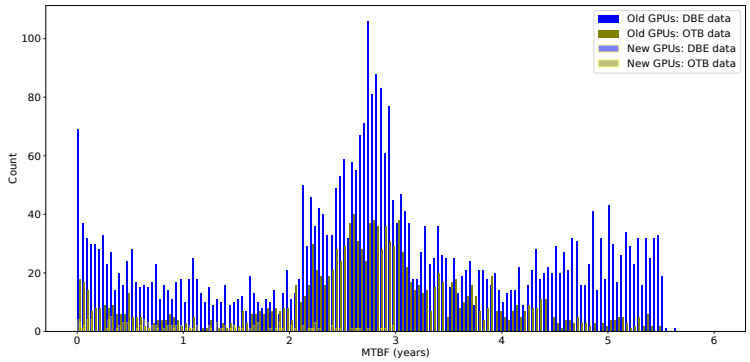
U.S. DEPARTMENT OF ENERGY

2014: Reliability of Titan is Focused on MTBF

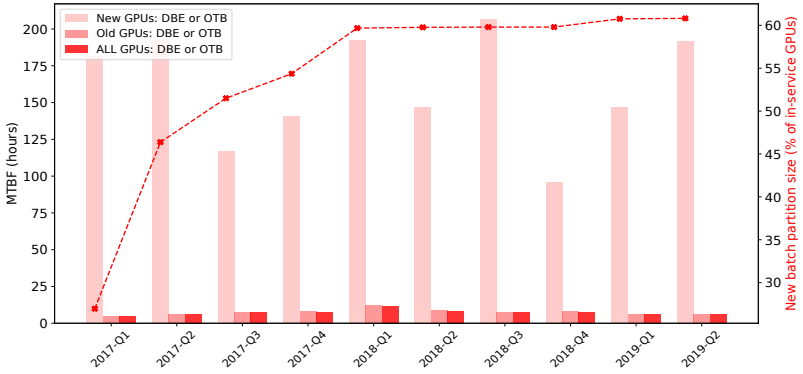
In 2015, we started seeing large numbers of GPU failures



System-wide Reliability: Quarterly number of failures (top) and MTBF (bottom).



Individual GPU Reliability: MTBF histogram for units that had at least one failure. Interpret carefully: lacks information from units with no failures!

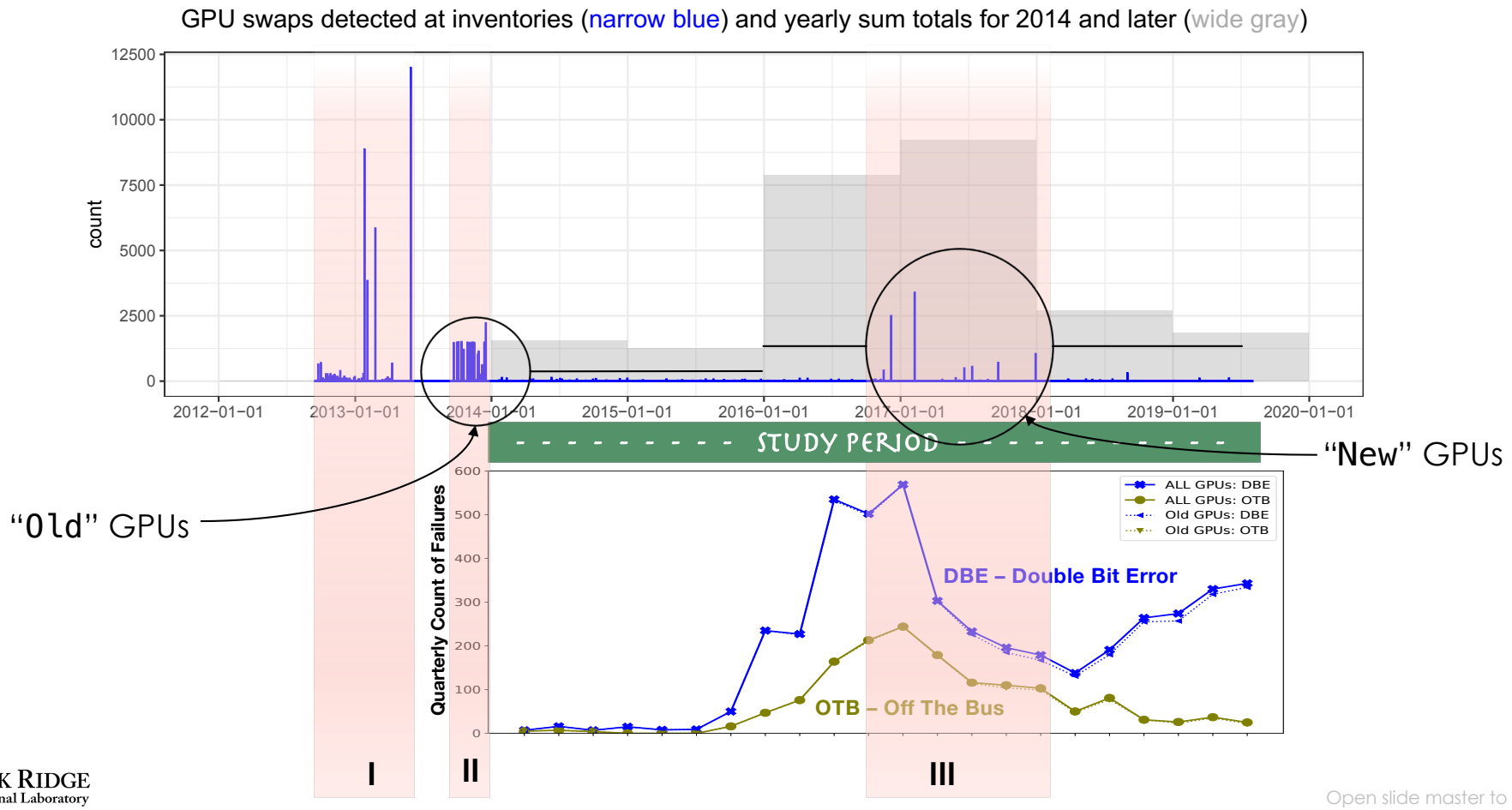


Old–New as Two Partitions: MTBF differs by 12x factor!

OAK RIDGE National Laboratory

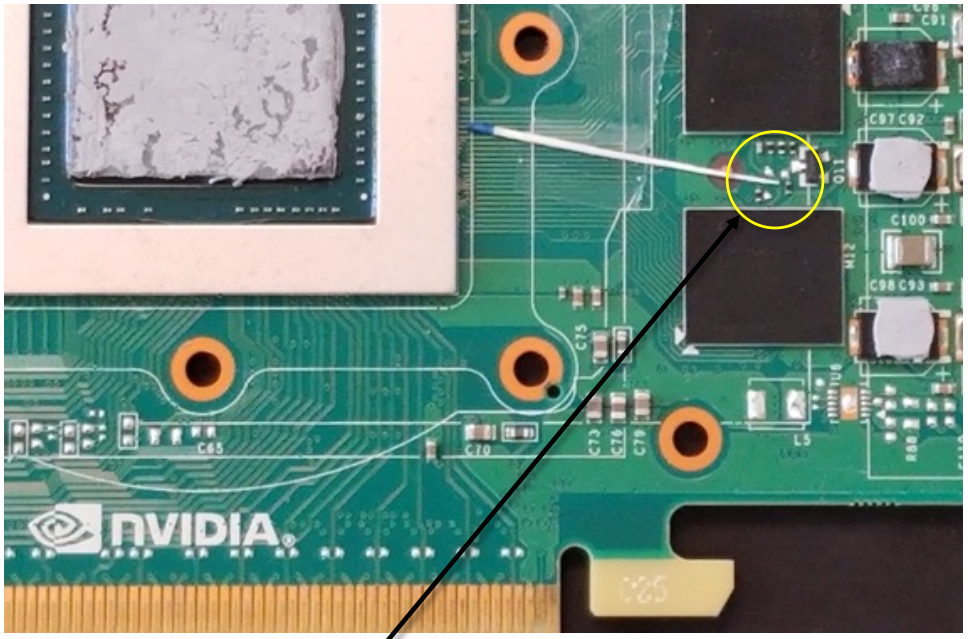
Ostrouchov, George & Maxwell, Don & Ashraf, Rizwan & Engelmann, Christian & Shankar, Mallikarjun & Rogers, Jim. (2020). GPU Lifetimes on Titan Supercomputer: Survival Analysis and Reliability. 1-14. 10.1109/SC41405.2020.00045.

Three Rework Cycles and Years of Stable Operation

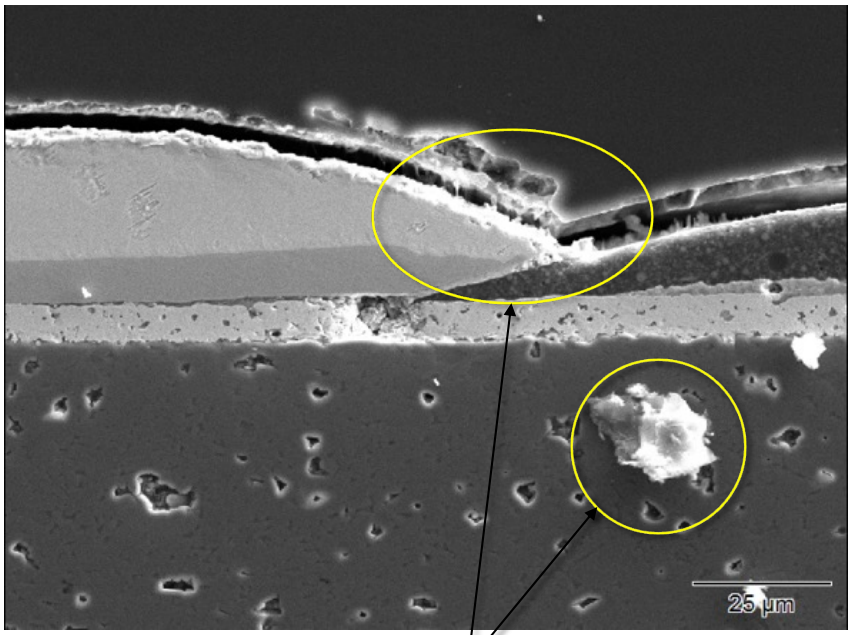


OAK RIDGE National Laboratory

Root Cause: Non-ASR Components on SXM GPU



NVIDIA SXM – Location of a non-ASR



Silver-sulfide corrosion
"Flowers-of-Sulfur"

ASR = Anti-Sulfur Resistor

We had logs of hardware swaps and failures, but without well-established analytics tools it took us **much** longer than we would have liked to diagnose and fix these issues

Data Platform Motivations

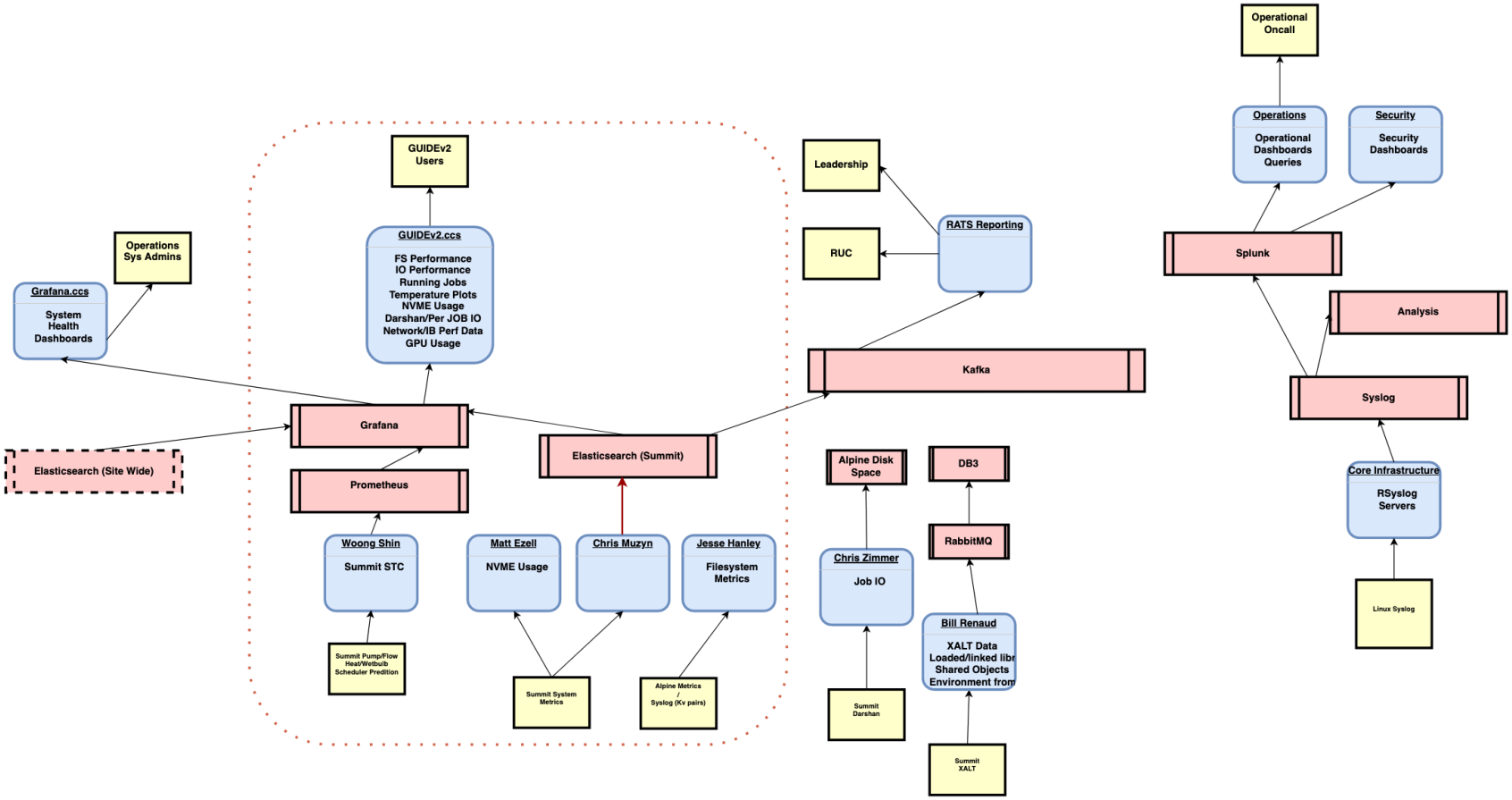
'Scaling' was no longer scaling

- Increased system complexity and changes to both systems and schemas over time made data analytics incredibly manual
- We needed to replace batch processing and enable stream processing where possible
- Traditional data sinks did not provide flexibility of modern data warehouses for applications users wanted to use
- A central data bus was necessary to decouple opaque data pipeline sources and sinks and provide O(n) scaling

New technologies made this possible

- Several scalable, robust, flexible message busses were becoming mature
- Modern data warehouse designs and search/analytics tools like Elastic were being explored by various teams
- Data analytics tools were maturing and our operations teams were becoming more capable of slicing and dicing telemetry streams
- Platform as a service (PaaS) had just been deployed within NCCS and was reducing administrative burden

2019: NCCS Analytics and Monitoring 'Platform'



OAK RIDGE National Laboratory

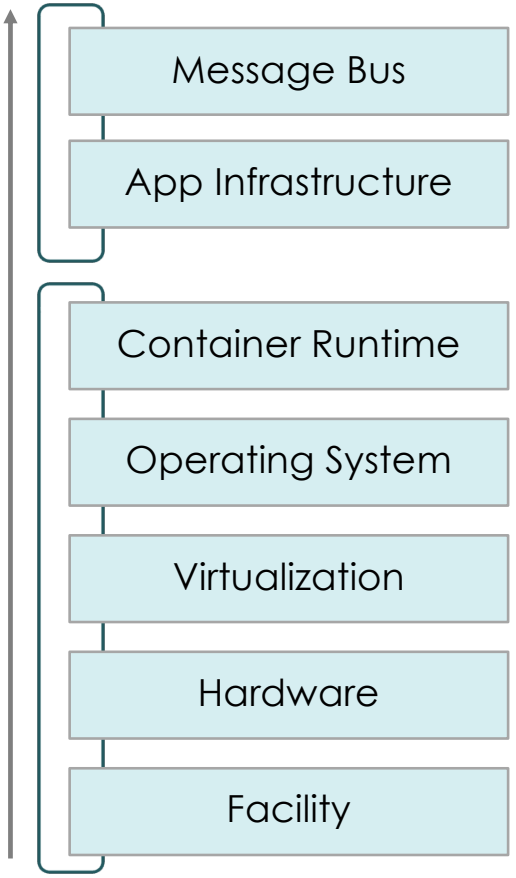
Data Platform Strategy – We needed a new capability

What is the scope for this new capability?

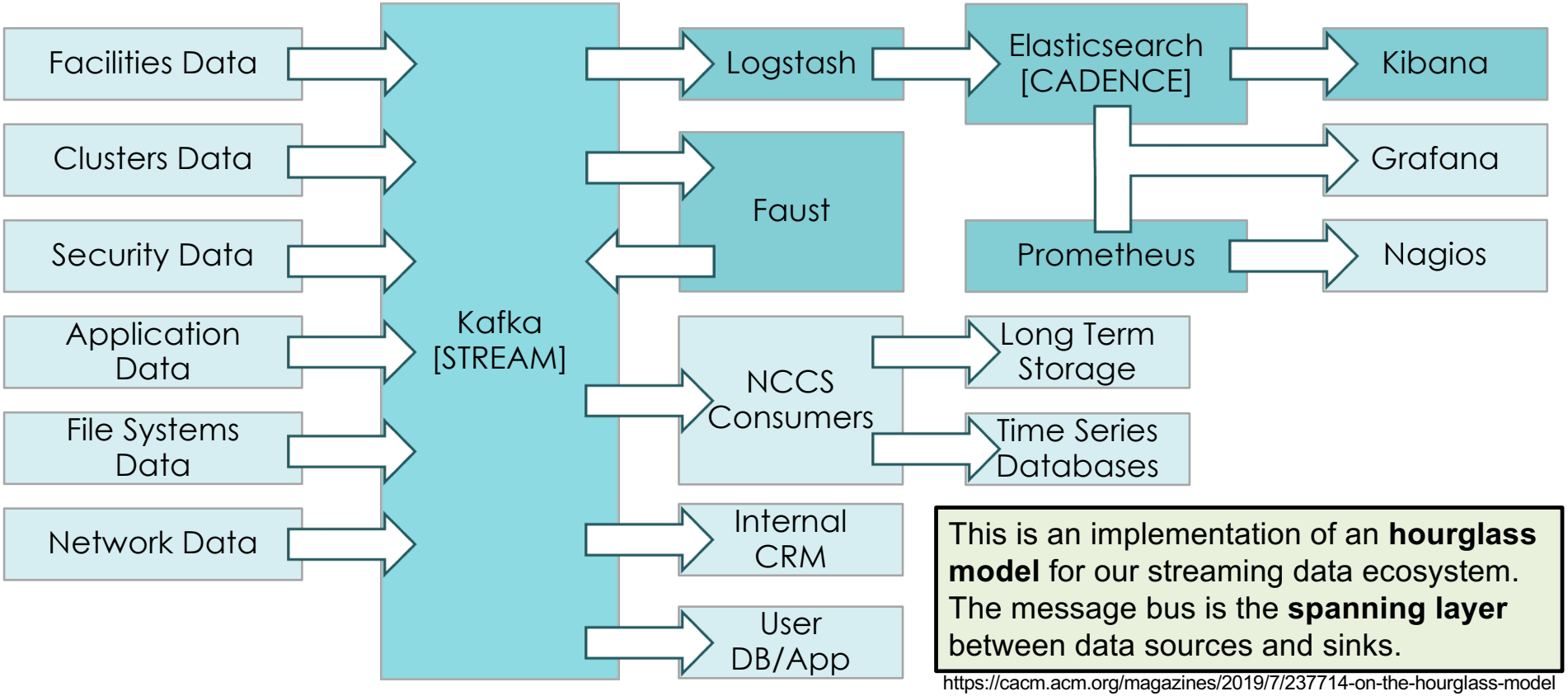
- Use operational / SRE best practices to provide data assurance
- Reduce the 'data wrangling' that scientific end users have to do
- Be advocates for both data producers and consumers
- Inform institutional data policy and help resolve data ownership conflicts

What is the application stack?

- Deliberately focus on 'top of stack' to support applications, platform users, and data analysis
- Utilize PaaS for supporting layers to minimize complexity
- Cleanly and clearly define data pipeline roles and responsibilities between consumers and producers



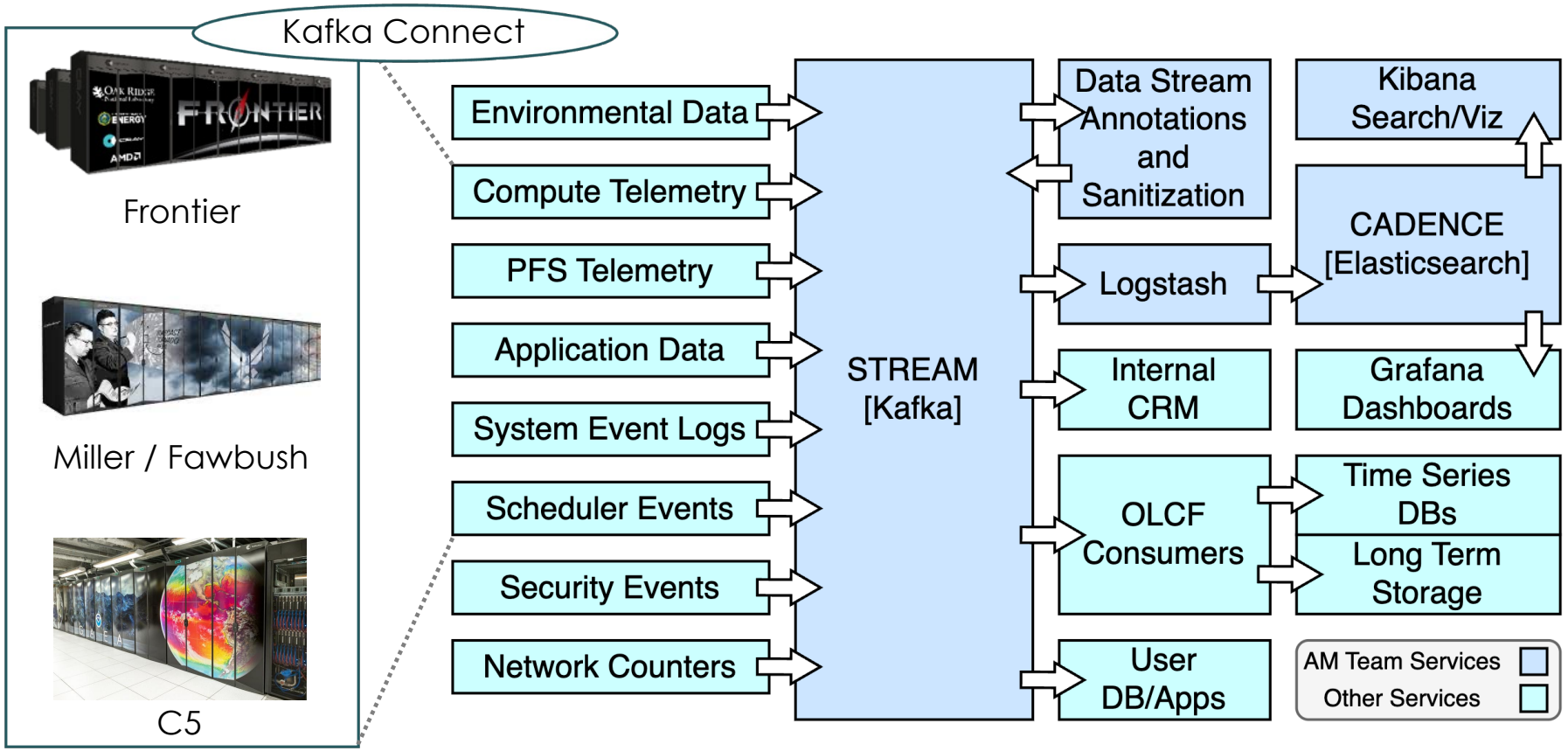
2021: NCCS Analytics and Monitoring Platform



This is an implementation of an **hourglass model** for our streaming data ecosystem. The message bus is the **spanning layer** between data sources and sinks.

https://cacm.acm.org/magazines/2019/7/237714-on-the-hourglass-model

STREAM Architecture in 2023



Data Platform Challenges

Sustainability

- Data sources ***will change*** over time
- Systems will come and go and technology will change
- Technical debt can be difficult to reduce once accrued
- Once automation exists for production and consumption… good luck!

Documentation

- Data producers should, ***in theory***, be the best equipped to answer questions about data sources
- Data consumers typically ***don't have enough context*** to understand the information they receive through telemetry pipelines

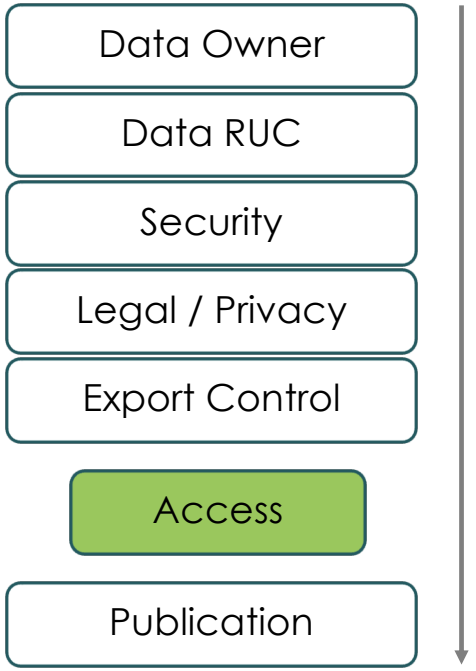
Performance / Robustness

- Controlling types and sizes of data can be challenging – ***data throughput tends to grow over time***
- Monitoring individual topics can be difficult, especially when ***a few key topics dominate*** systems engineer time

Lessons Learned – Access Control

Controlling Access

- We define a data 'owner' to be the producer of data, and we give some control over who can access streaming messages.
- Data 'consumers' apply for access and are granted individual topic credentials based on need.
- The OLCF has an interest in reviewing potential research outcomes and discoveries
  - Misinterpretation of information is quite common!



Lessons Learned – Topic Naming

Topic Naming

- Changes to topic names as well as changes to client configuration is very difficult to manage
- We developed a sustainable topic naming scheme based on use cases
- NCCS uses a delimited topic name 'tuple' based on data source owner, system name, the subsystem that produces messages, and the specific topic subject the topic is about
- Example:

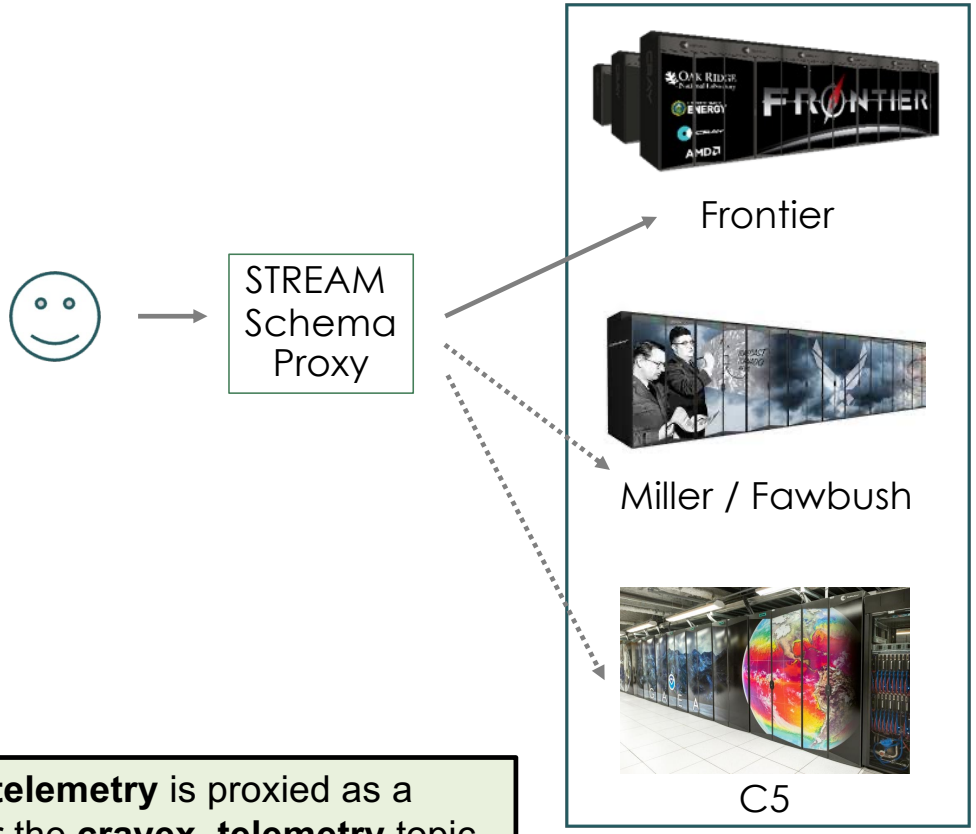
stf002hpc.frontier.hpcm.crayex_telemetry

Source System	Subsystem	Topic Subject
frontier	hpcm	HPCMLOG
c5		SYSLOG
t5		crayex_alerts
miller		crayex_telemetry
fawbush		event_cooldev
ace		hpcm_inventory
		hpcm_inventory_dimm
		log_iml
		powerservice_operations
		powerservice_rawpower
		sensors_node
		slurm_jobs
		…

Lessons Learned – Schema Registries

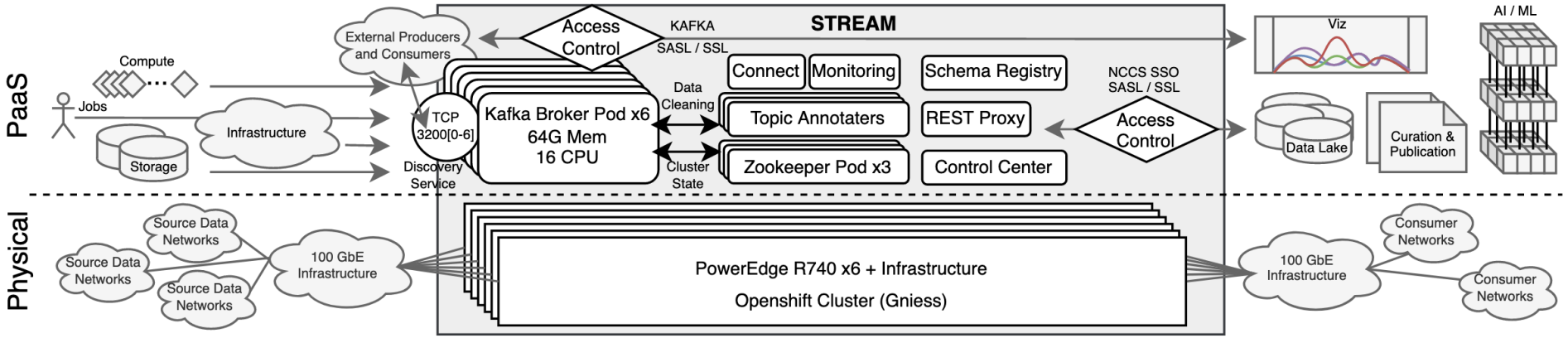
There are many, many registries!

- HPCM Kafka schema registries on various systems may not be configured in the same way
  - The ordering of topics and versioning of topics over time lead to different schema definitions for the 'same topic' across systems
- OLCF developed a fairly simple flask application to 'proxy' schema registry access
  - On client access to STREAM, schema registry request is modified to connect to HPCM schema registry of the system the topic is produced from



A user request for **stf002.frontier.hpcm.crayex_telemetry** is proxied as a request to the Frontier schema registry service for the **crayex_telemetry** topic.

Lessons Learned – Monitoring and Break/Fix



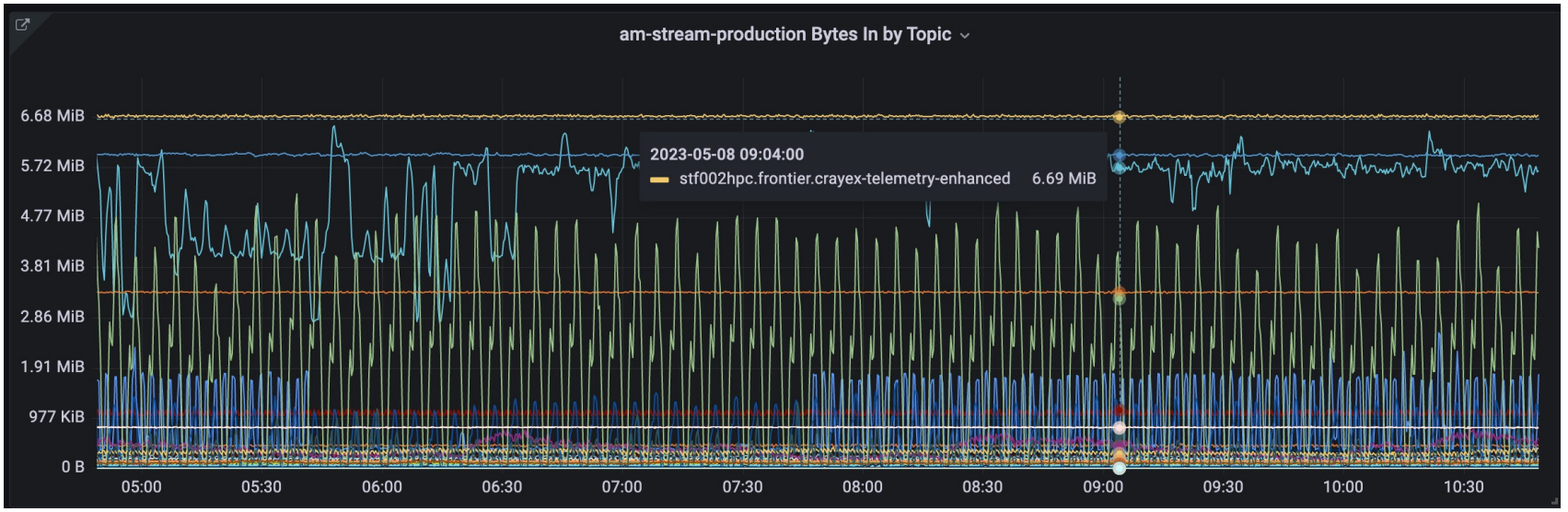
There are many layers of support: from HPE engineers to systems admins to end user applications. Lots of arrows here mean lots of opportunity for breakage! We **monitor the monitoring tools** as best as we can and integrate alerting with existing systems. Buffering at each layer can help preserve data while troubleshooting occurs.

Characterization of crayex_telemetry on several systems

Topic name	Status	Partitions	Production (last 5 mins)	Consumption (last 5 mins)	Followers	Observers	Last produced
stf002hpc.ace.hpcm.crayex_telemetry	Healthy --	1	27.95KB/s	98B/s	3	0	May 8 2023, 11:1
stf002hpc.c5.hpcm.crayex_telemetry	Healthy --	7	1.05MB/s	786.03KB/s	21	0	May 8 2023, 11:1
stf002hpc.fawbush.hpcm.crayex_telemetr	Healthy --	7	434.49KB/s	443.88KB/s	21	0	May 8 2023, 11:1
stf002hpc.frontier.hpcm.crayex_telemetry	Healthy --	40	5.95MB/s	6.75MB/s	120	0	May 8 2023, 11:1
stf002hpc.miller.hpcm.crayex_telemetry	Healthy --	1	169.29KB/s	147.76KB/s	3	0	May 8 2023, 11:1
stf002hpc.t5.hpcm.crayex_telemetry	Healthy --	7	19.03KB/s	19.03KB/s	21	0	May 8 2023, 11:1

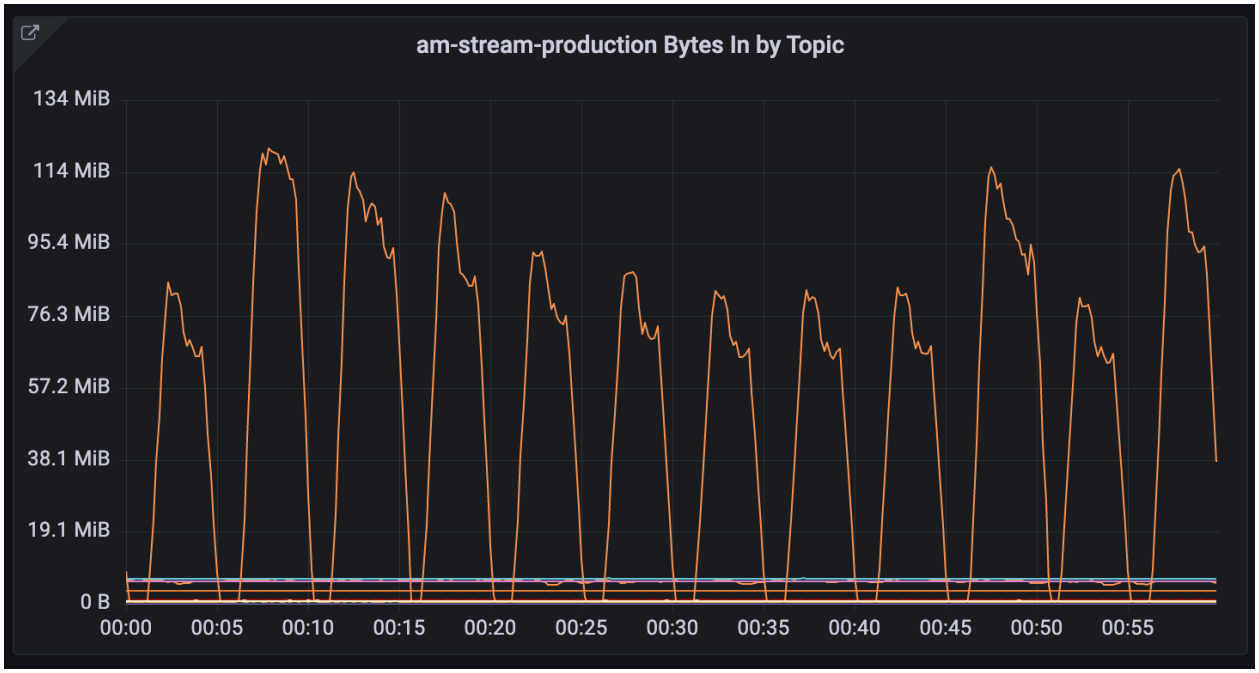
Corresponding topics for **crayex_telemetry** from Ace, C5, Fawbush, Frontier, Miller, and T5. Note the differences in partition count, production, consumption, and followers.

Selected Examples of Interesting Behavior



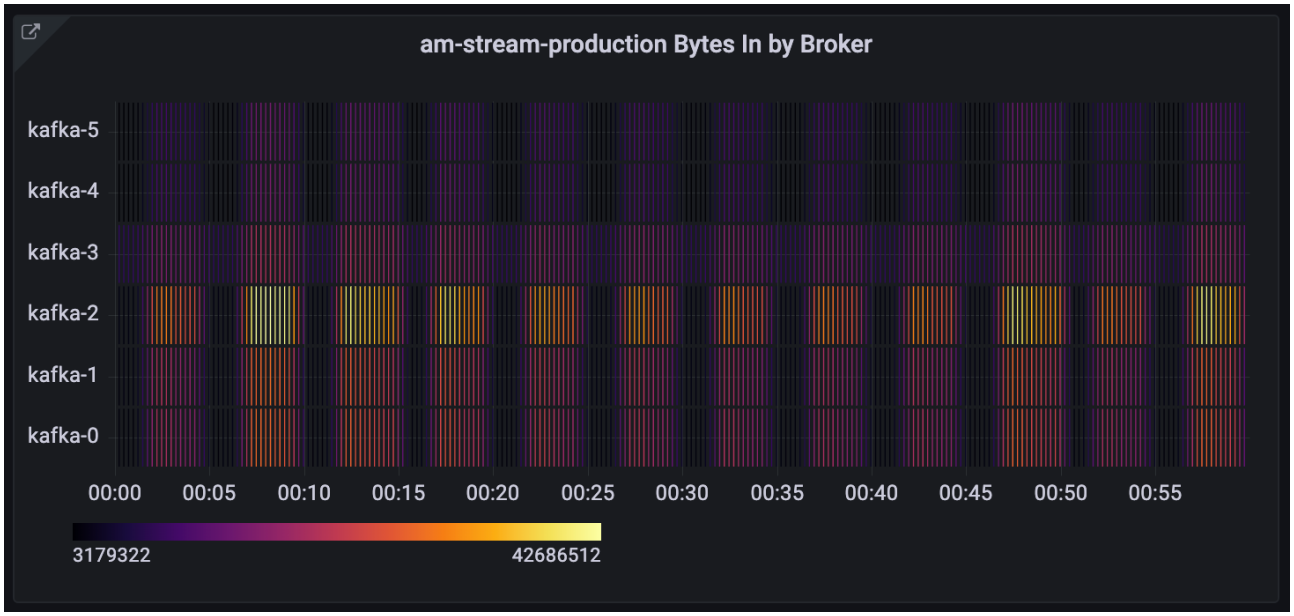
Typical 6-hour window of ‘steady state’ streaming telemetry performance. Most of STREAM’s 223 topics are < 1MiB/s and are dominated by several large topics.

Selected Examples of Interesting Behavior



A 1 hour window showing 5-minute collection periodicity of Lustre trace data for Orion.

Selected Examples of Interesting Behavior



Monitoring bytes in per broker with a heatmap diagram reveals periodicity as well as an unbalanced amount of information flowing to the kafka-2 broker for this time interval.

Discussion

Acknowledgements:
This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.