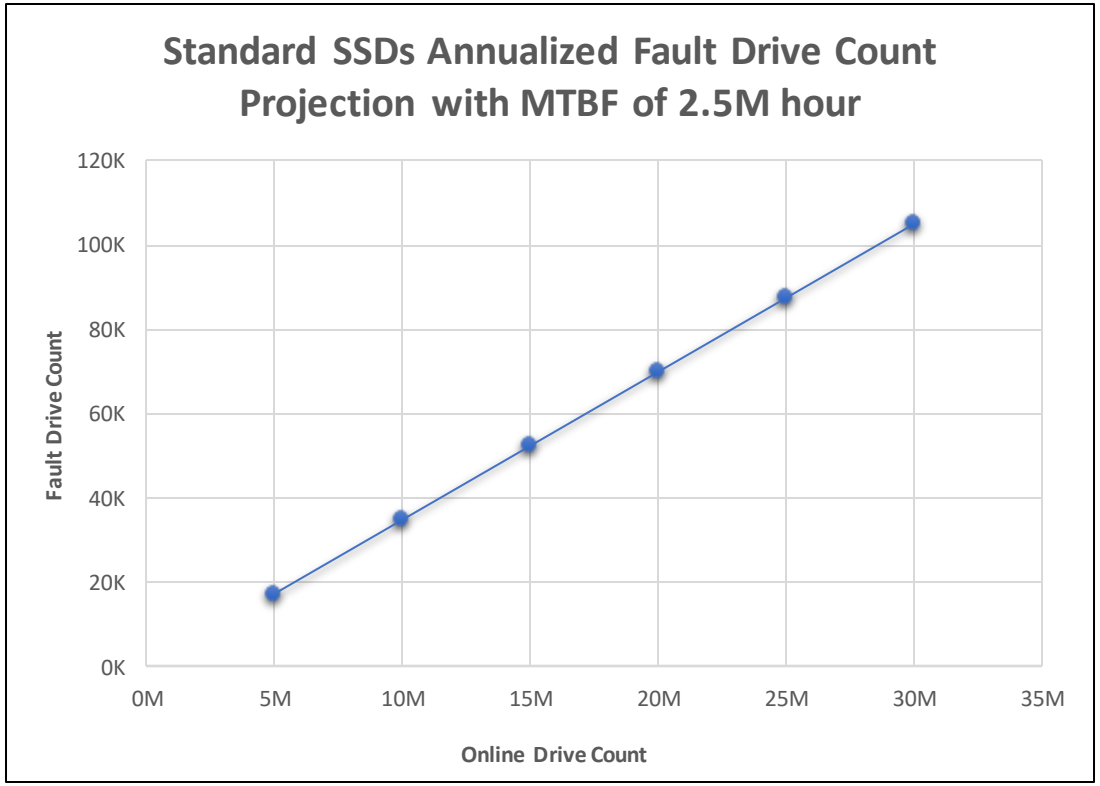
Flash Memory Summit

Vertically Integrated High Resilience SSDs Designed for Cloud Computing

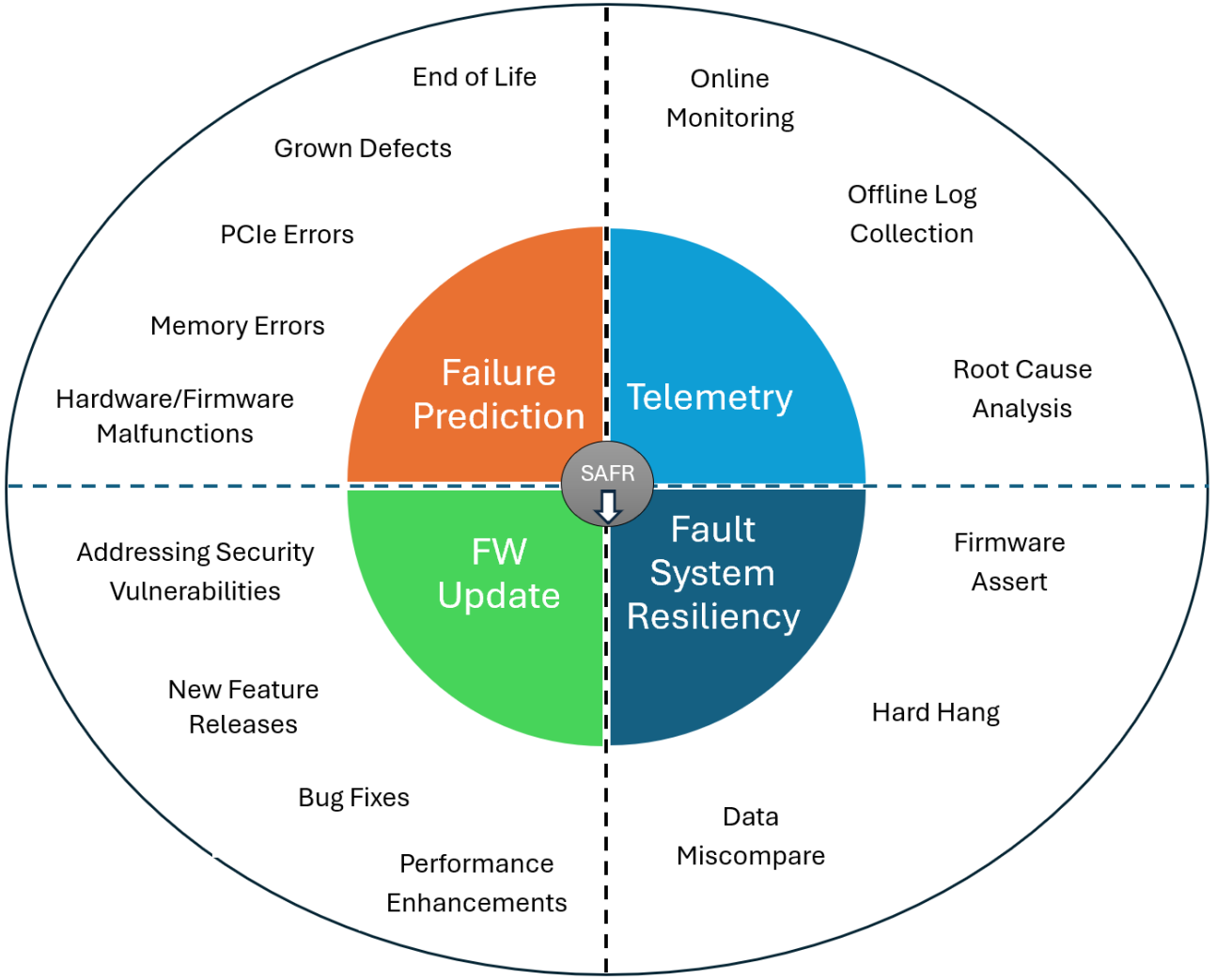
Presenter: Ayberk Ozturk, Microsoft

Objective

- There exist tens of millions of Solid-State Drives (SSDs) within the Cloud infrastructure.
- Given the standardized SSD Mean Time Between Failures (MTBF) of 2.5M hour or Annualized Failure Rate (AFR) of 0.35% prevalent in the industry, a significant number of storage devices, in the hundreds daily or tens of thousands yearly, encounter failures.
- To address this challenge, Microsoft is proactively engaged in enhancing the quality of storage devices, minimizing capacity impact, and improving customer experience through **vertically integrated high resilience (VIHR)** SSDs as specified in OCP v2.5 and above.
- VIHR reduces overall system annualized failure rate (SAFR) significantly.



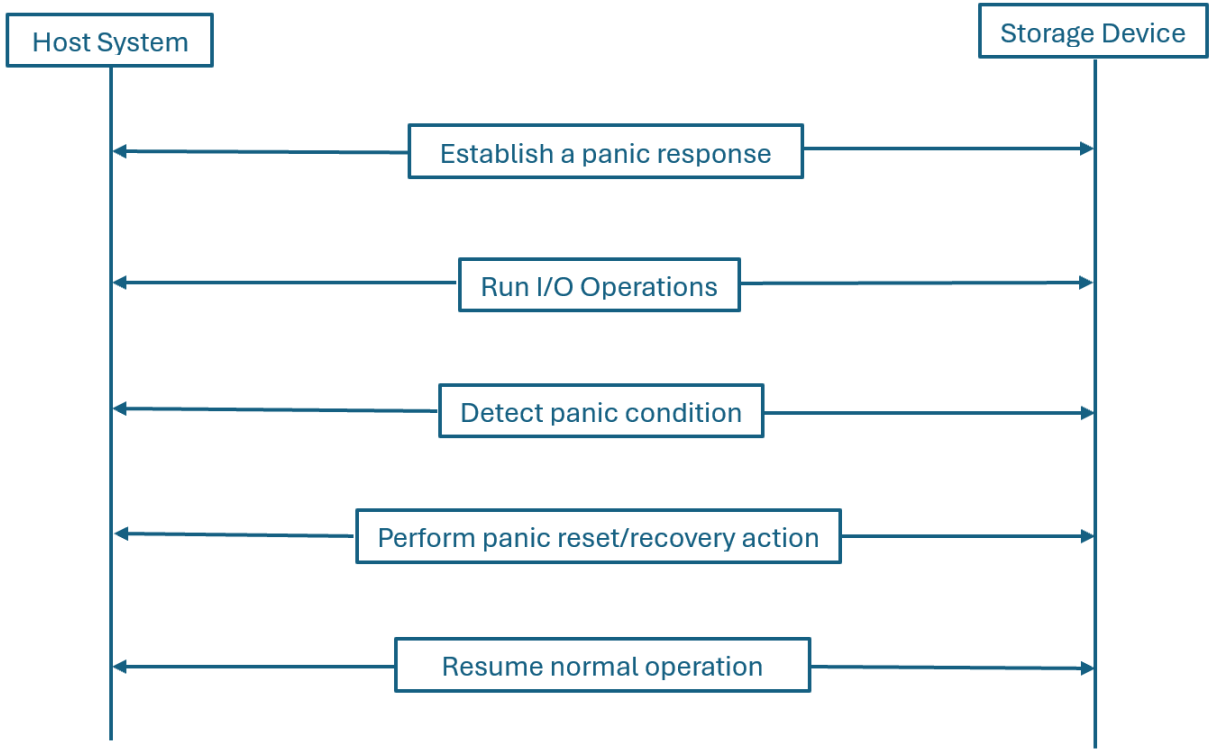
Key Capabilities and Applications of VIHR SSDs



Facilitating these four essential capabilities effectively mitigates the system-level impact of SSD failures.

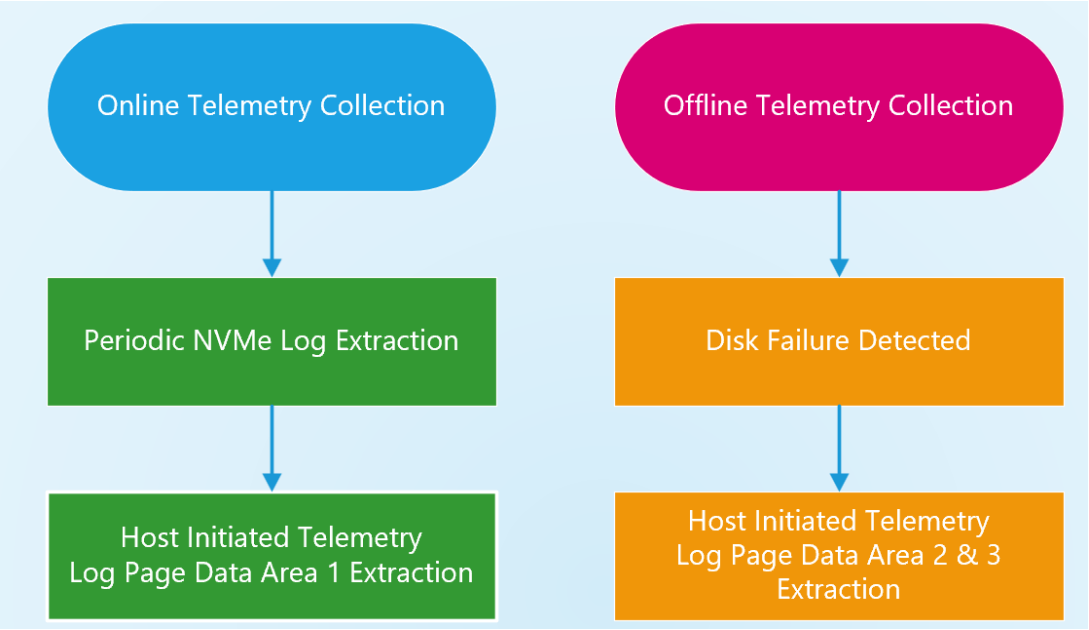
Fault System Resiliency

- Panic is a firmware assert, crash or hard hang.
- This Fault System Resiliency (FSR) capability detects and recovers from a panic on a storage device.
- A panic response is established through 0xC1 Error Recovery Log Page and Asynchronous Event Request (AER).
- A panic condition is identified through CSTS.CFS bit or AEN.
- Storage device performs one or more non-invasive corrective actions to recover from a detected panic condition to facilitate a normal operational state between a host system and the storage system.
- Systems with panic drives are moved to offline gracefully without impacting workload for further inspection and repair.



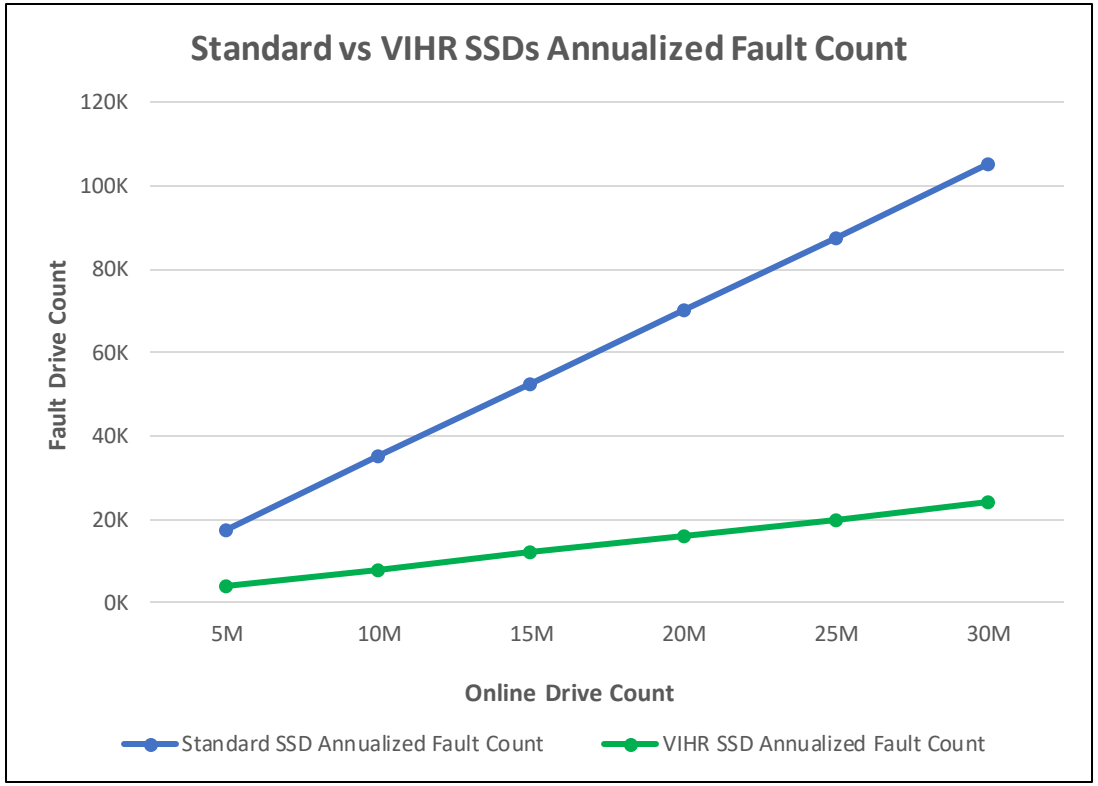
Telemetry Collection

- Standardized host and controller-initiated telemetry log page including vendor specific debug data section.
- NVMe Logs including Telemetry Log Page Data Area 1 are collected regularly from online drives to monitor devices health and predict failures.
- Vendor specific debug data including Data Area 2 & 3 of Telemetry Log Page is collected at the time of failures for root cause analysis.
- Telemetry attributes are specified to root cause all intermittent failures including, data miscompare, IO timeout and IO errors or latency spikes along with persistent failures.



Results

- Vertically Integrated High Resilience (VIHR) SSDs with 4 key capabilities of Telemetry, FW Update, Failure Prediction and Fault System Resiliency have resulted in a reduction of:
  - The system-level annualized failure rate of SSDs to 0.09%
  - Concurrently elevating the Mean Time Between Failures (MTBF) to 10 million hours
- The benefits of system failure savings increase as the fleet grows.



Conclusion

- The utilization of Vertically Integrated High Resilience (VIHR) SSDs presents a cost-effective strategy for constructing fault-tolerant systems.

- Continuous improvements to achieve North Star goal of 100% system resiliency:
  - Increasing the fault system resiliency coverage will detect and handle more fatal failures gracefully.
  - Enhanced telemetry with new critical drive attributes will root cause more failures and improve failure prediction further.