Enhancing SSD Diagnostics for Hyperscale Deployments

Venkat Ramesh

Production Systems Engineer, Meta

FMS the Future of Memory and Storage

Agenda

1

Role of Diagnostics for Hyperscalers

2

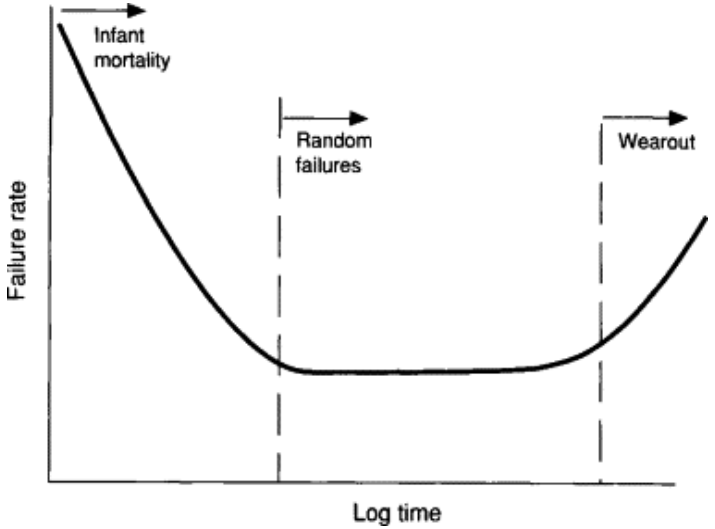
Case Study: Enhancing Diagnostics for Accelerated Life Testing

3

The Road Ahead

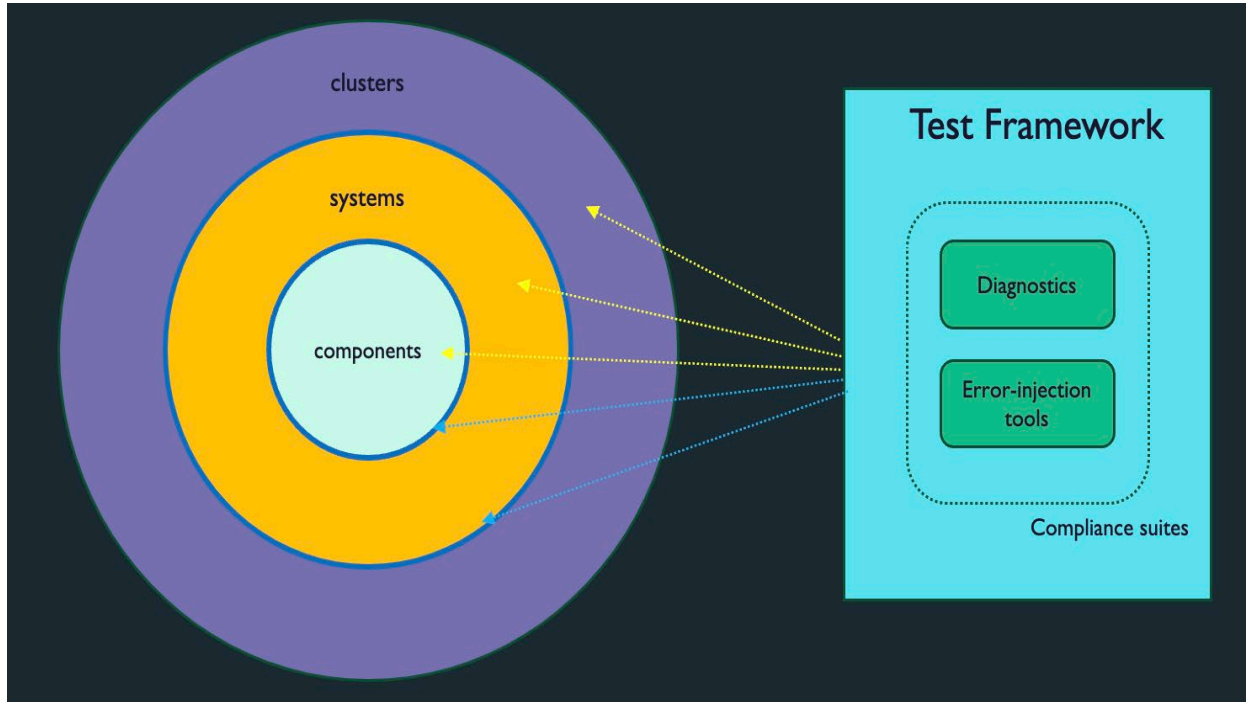
Catering to Evolving SSD Health & Environments

- Non-linearity in failure rates and failure modes
  - Wear manifests as Latency, IO Errors
  - Unearthing of latent test escapes
- Applicability of early-stage tests, reproducibility at scale are paramount
- Diagnostics can serve as feedback mechanisms between suppliers and hyper-scalers, from design and production.

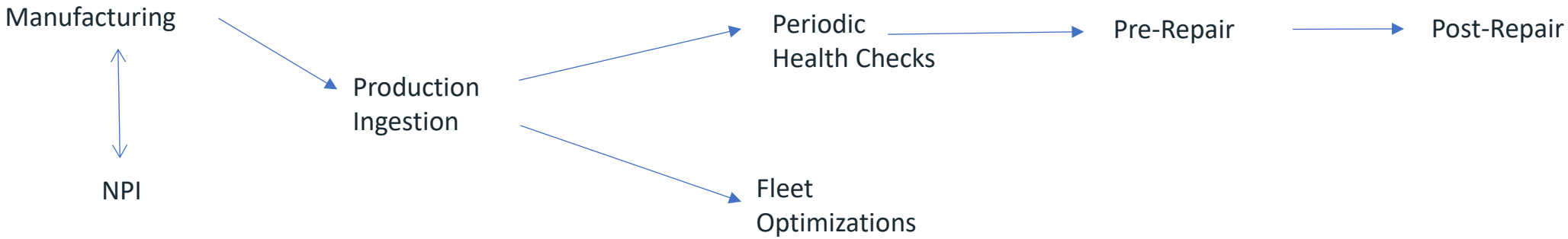


Kernel of a Diagnostic

- OCP Test & Validation Initiative: Provides evaluation of compliance for DUT, with evidence to support
- Integrates seamlessly with variety of Test Executives across deployment environments and scale
- Interop with other component diagnostics to provide cluster-level health view through intrusive testing.



Hyperscale Use-Cases and Deployment Environments



- Varying Constraints
  - Runtime (few hours/days in NPI to few seconds for periodic checks in fleet)
  - Intrusiveness-levels (workload QoS impact to induced wear)
  - Workload diversity: Span four-corner stress as well as target application emulation
- Data-schema dimensionality should capture environment metadata to correlate behavior across stages.
- In AI training clusters, uptime/fault-isolation requirements are very strict

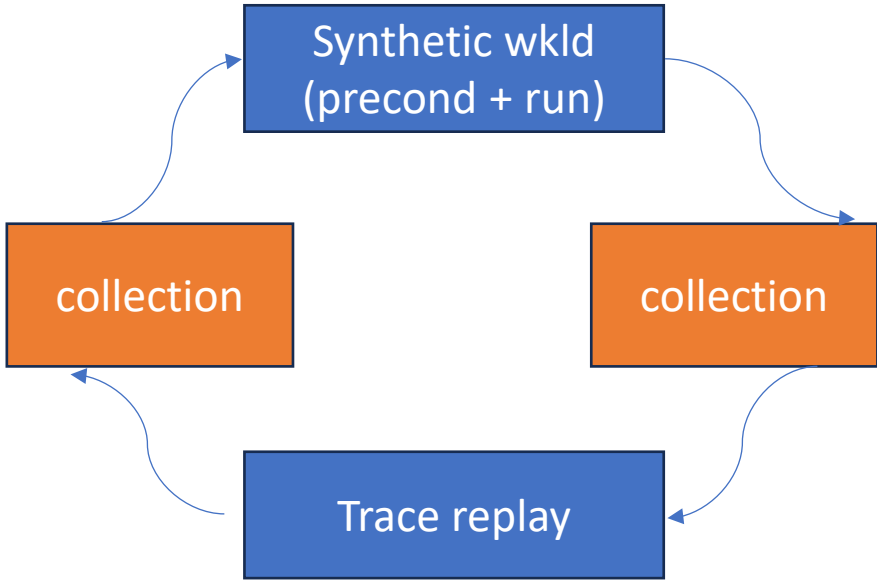
Case Study: Accelerated Life Testing Diagnostic Needs

TestDrive

- Goals:
  - Measure and bound acceleration factor computation to take a drive from point A to B in its life-cycle with a given workload at an acceptable acceleration
  - Stop to capture any failure for repro/root-cause
- Use-cases:
  - Supplier RDT during qual phase
  - Empirical confirmation of fall-out rates/fixes/optimizations
  - Forecasting & repair planning
- Existing Landscape:
  - Supplier-provided tools with custom formats
  - Private channels for workload sharing
  - Missing continuity of diagnostic data across HW Lifecycle

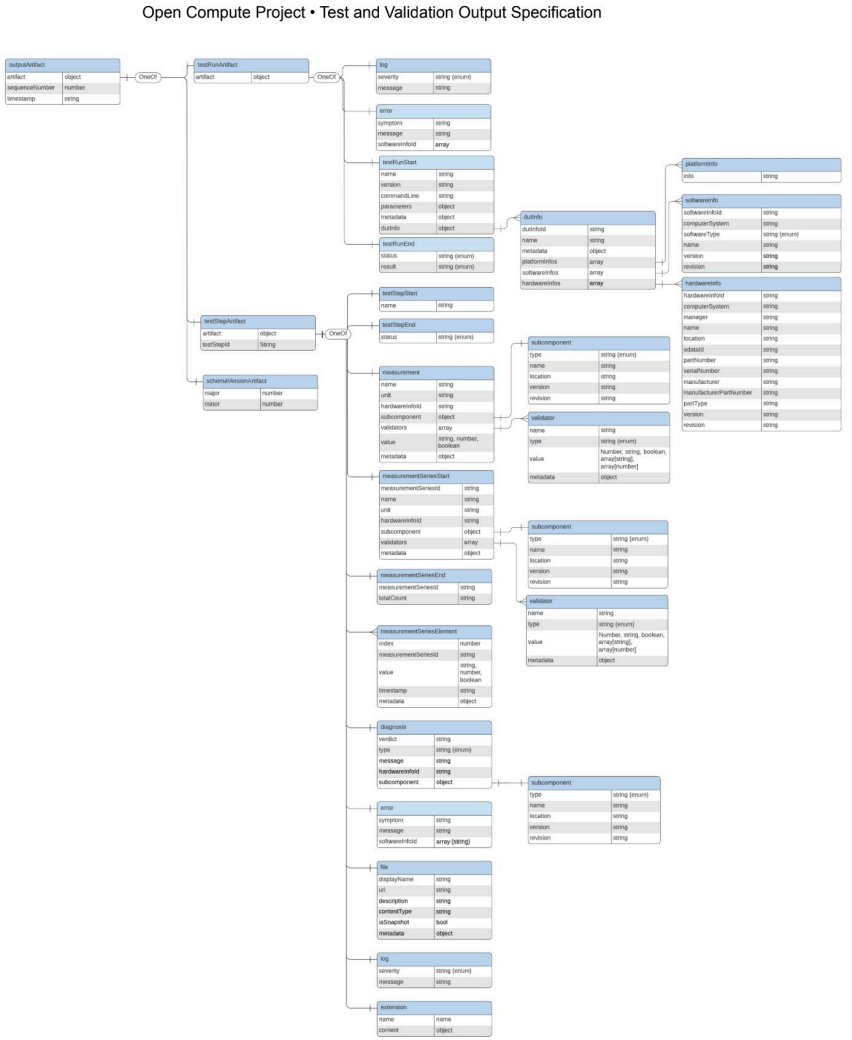
TestDrive: Workloads & Heuristic

- Intermingle synthetic emulations and workload traces
  - FioSynth repository from Meta
  - Live workload traces
- Periodic collection of SMART telemetry, perf and error counters
  - Statistical and threshold-based pass/fail
- Real-time detection of latency and kernel errors
  - Debuggability on failure is huge concern
- Knobs for plugging in rated-endurance source of truth and tune runtime/ rate-limit intrusiveness



Case Study: Reporting & Analytics

- Leveraging OCP T&V data model to unify across various executors and capture hardware, platform and software Metadata needed for diagnosability.
- Can correlate with behavior in fleet to tune, predict QoS behavior and failure trends
- Provide empirical evidence to support capacity reallocations, swap rate tuning.



The Future: A Comprehensive Diagnostics Package for OCP-HW

- OCP T&V community is building up a community of diags for Open Systems for AI: for CPU, Memory, LMT etc
- For Flash, OCP compliant SSDs have standardized telemetry collection enabled by nvme-cli OCP plugin and storage tests have been contributed by Hyper-scalers
- Meta committed to bring more capabilities such as Accelerated Lifecycle Tests and align behind OCP T&V
- Hyper-scalers bring workloads, suppliers can help standardize methodology and reporting

Let's go on a drive!

- Join OCP T&V workgroup to learn more, adopt and contribute to ocp-diags.
- Stay tuned for ocp-diags from Meta for Flash in Open Compute Project GitHub.
- Reach out to share feedback and ideas.

Thank You!