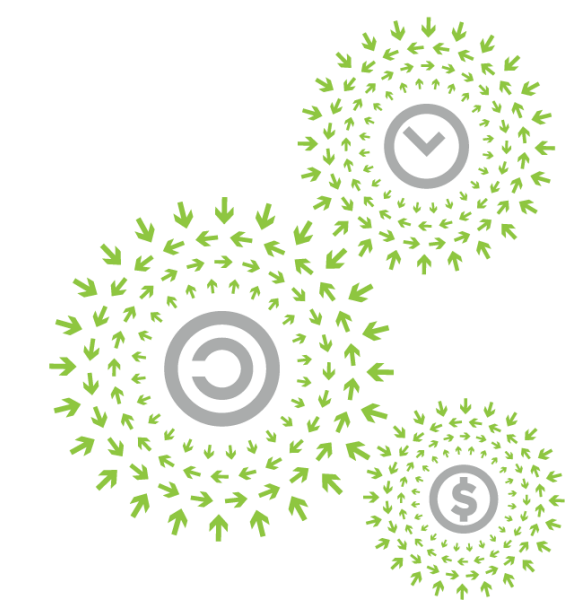# Open Cloud Server Management in the Exchange Online Service

Matt McCormick
Senior Program Manager, Microsoft

Open. Together.

Agenda

Workshop goals:

Share our OCP "customer" perspective on managing Open Cloud Server (OCS) and Project Olympus hardware at scale and discuss opportunities for the Hardware Management group to improve this experience.
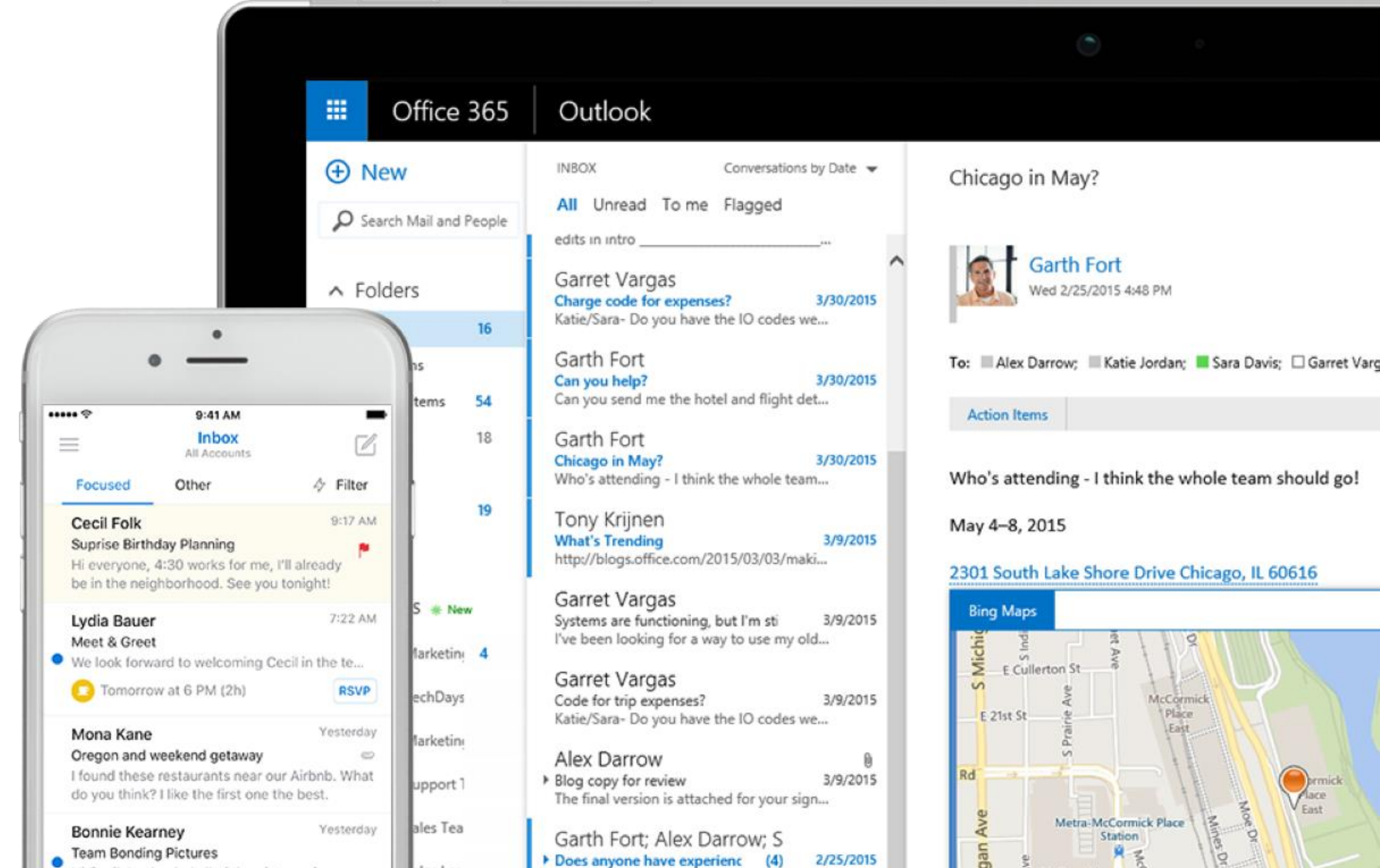
Workshop benefits:

Learn about Microsoft "first-party" deployment of OCP servers across a key service offering, how we manage this fleet, and how we've engineered solutions to current manageability limitations.

Open. Together.

Exchange Online Service Background

Exchange Online

Business-class email and calendaring for enterprise

Effortless email and calendar for consumers
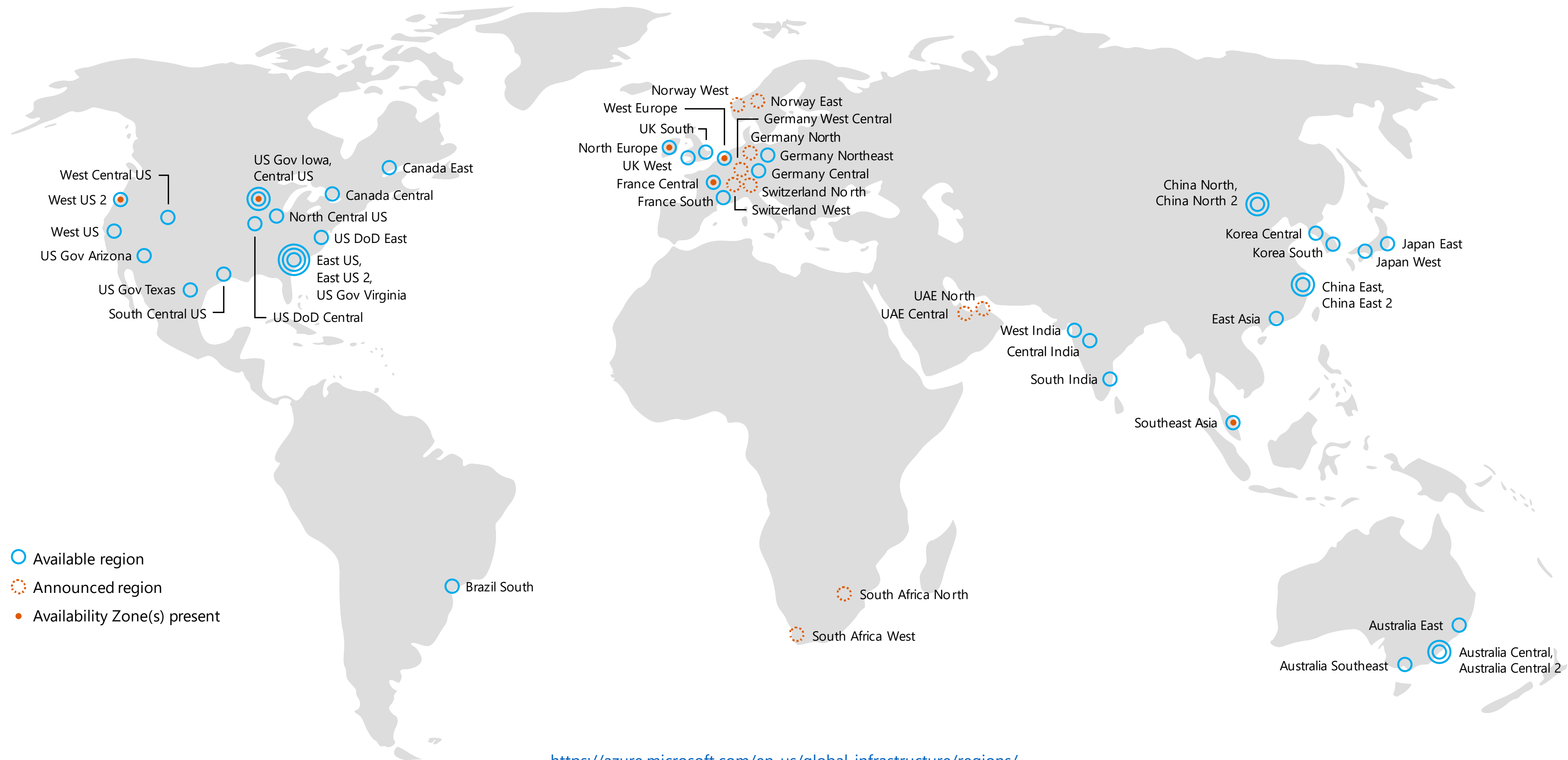
Exchange Online by the Numbers

- 100+ datacenters

- 10s of 1,000s of racks

- 100s of 1,000s of servers

- 1,000,000s of disks

- Exabytes of storage

Microsoft Datacenter Regions

https://azure.microsoft.com/en-us/global-infrastructure/regions/

OCP Hardware Usage in the Exchange Online Service

Open Cloud Server (OCS) Chassis Manager

The M1030 Chassis Manager is an ARM processor-based board for enabling management of Server Blades and a WCS Chassis. It is designed to act as a hot-swappable plug-in module for a WCS Chassis.
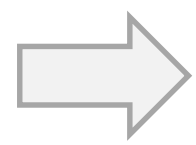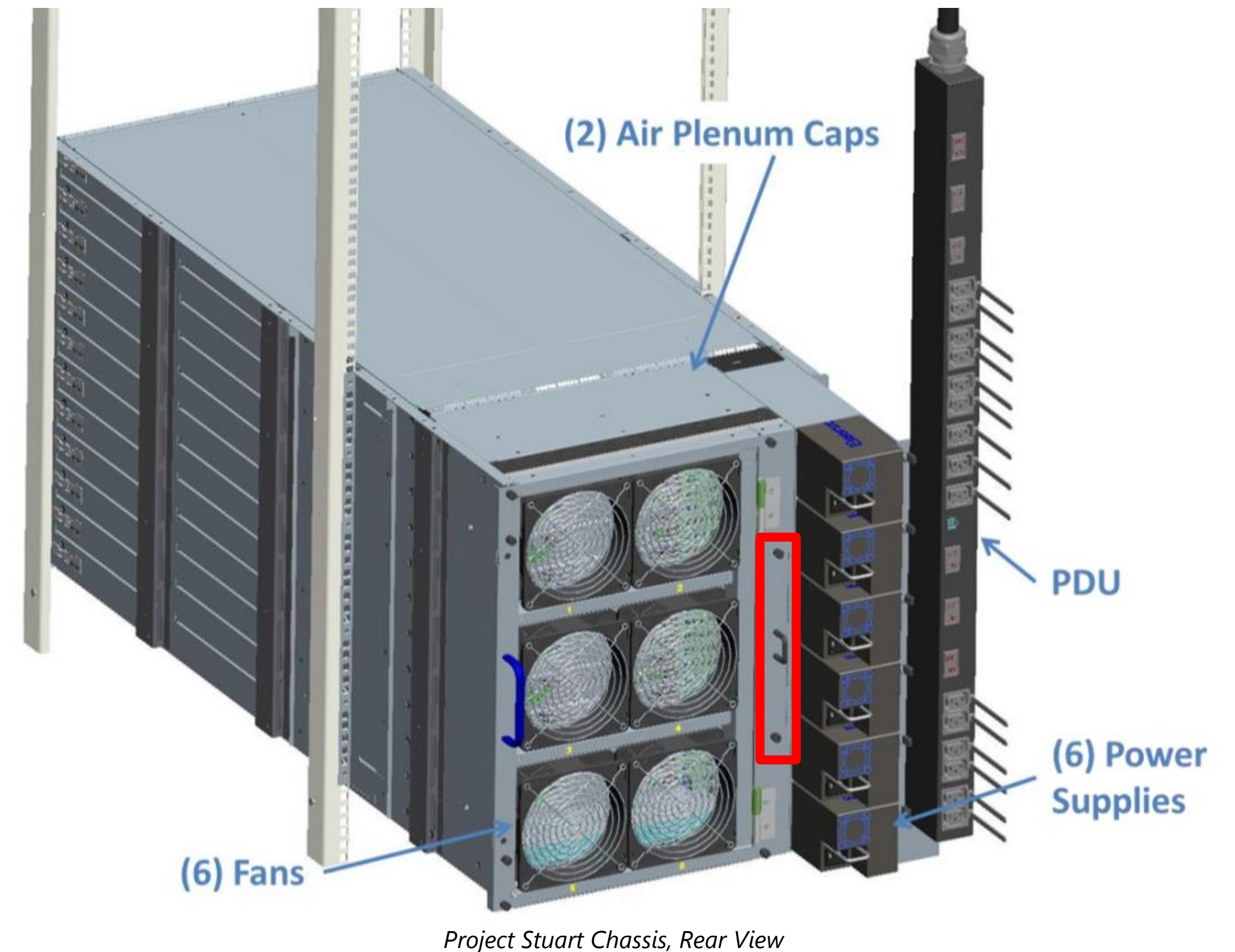
https://www.opencompute.org/wiki/Server/SpecsAndDesigns-old#Open_CloudServer

Specifications

(2) Air Plenum Caps

PDU

(6) Power Supplies

(6) Fans

Project Stuart Chassis, Rear View

Open. Together.

Project Olympus Rack Manager

The M2010 Rack Manager is an ARM processor-based board for enabling management of Server Blades and a WCS Rack. It is designed to act as a hot-swappable plug-in module for a WCS Power and Management Distribution Unit (PMDU) or as part of a separate Rack Management Module for supporting Non-WCS racks.

Specifications

https://www.opencompute.org/wiki/Server/ProjectOlympus
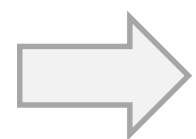
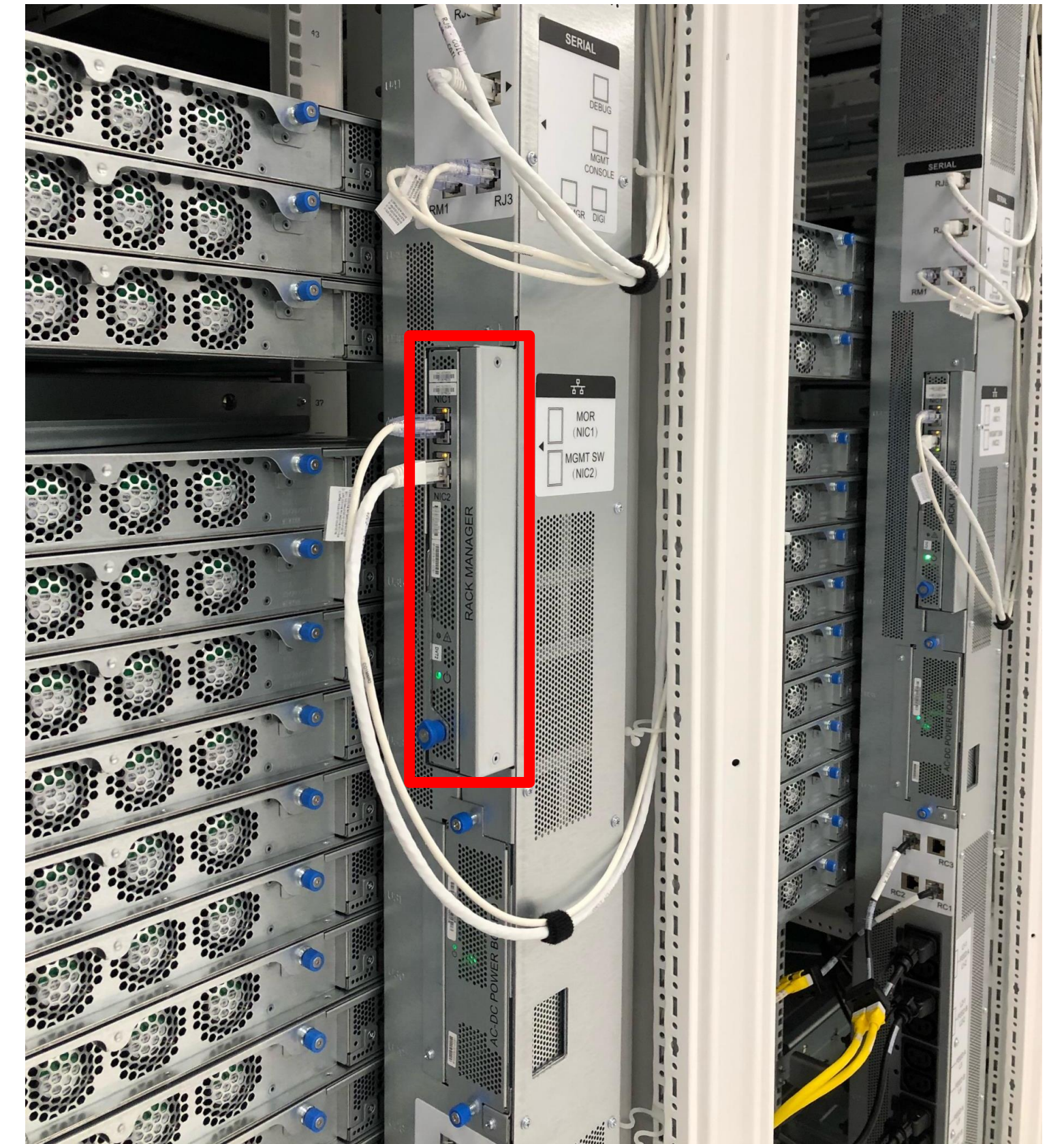
Project Olympus Compute Rack, Rear View

Open. Together.

Rack Manager Service

The Rack Manager service provides the front end through a Redfish-compliant REST API for automated management and a command-line interface for manual management. It manages all devices within the rack and communicates directly with the server management system through the network.

Embedded Software

https://github.com/opencomputeproject/Project_Olympus_rack_manager
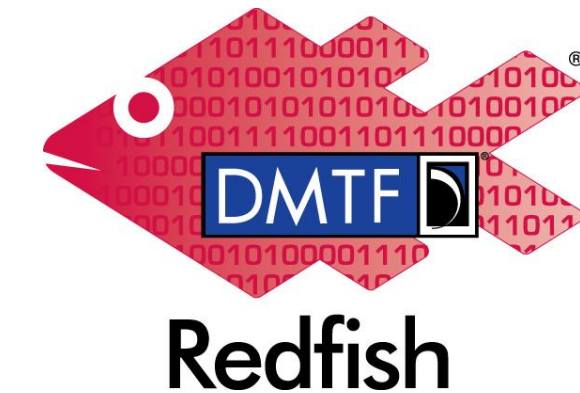
Redfish

User Management (add, remove, etc.)

PMDU Management (power state, relay state, meter alert, throttling, etc.)

Rack Manager Services (Redfish, TFTP, NFS, NTP, JTAG, etc.)

Rack Manager (rack inventory, firmware update, log services, attention LED, etc.)

System / Blade Management (power state, boot order, BIOS configuration, TPM status, mezzanine card status, remote media mount, etc.)

System / Blade Serial Sessions (session management, etc.)

Switch Management (port information, firmware update, etc.)

UPS Management (voltage, power, current, alarm, etc.)

High-level functionality exposed by Redfish OCP profile

Open. Together.

OCS Toolkit

The Open Cloud Server (OCS) Operations Toolkit is a collection of scripts and utilities for updating, diagnosing, and testing OCS servers.

Embedded Software

https://github.com/opencomputeproj ect/ocs-source-code-and-operations-toolkit-for-open-cloudserver

View-WcsConfig screenshot, OCS Toolkit

Hardware Failure Detection & Repair

BladeCommandFailure BuddyCableUptime CmBuddyNotFunctional CMInvalidCredential CMNetworkFailure CmNotRebooting CMRackFan CmServiceVersionCheck CmStuckInDeployment CompressionPerformance ControllerOverHeatingError ControllerOverHeatingWarning CriticalOSDisk DeviceAccessError DeviceOtherError DhcpScopeExhaustion DiskNonSwappableSSDFailure DnsListDecom DnsListEmpty DnsListOther DriverFirmwareMissing EventBugCheck EventCable EventController EventControllerBattery Event CacheAndBattery EventCpu EventDisk EventDiskPredictive EventLogFlighting EventMemory EventNIC EventServer EventWcsDisk F5BigIpCredFailure FlipDisk FpgaEventLog FpgaNotFunctional GatewayNotFound HardwareReserved Failure IbNandMotherboardFailure IncorrectNicMacAddress InvalidAssetData Inva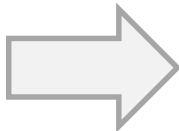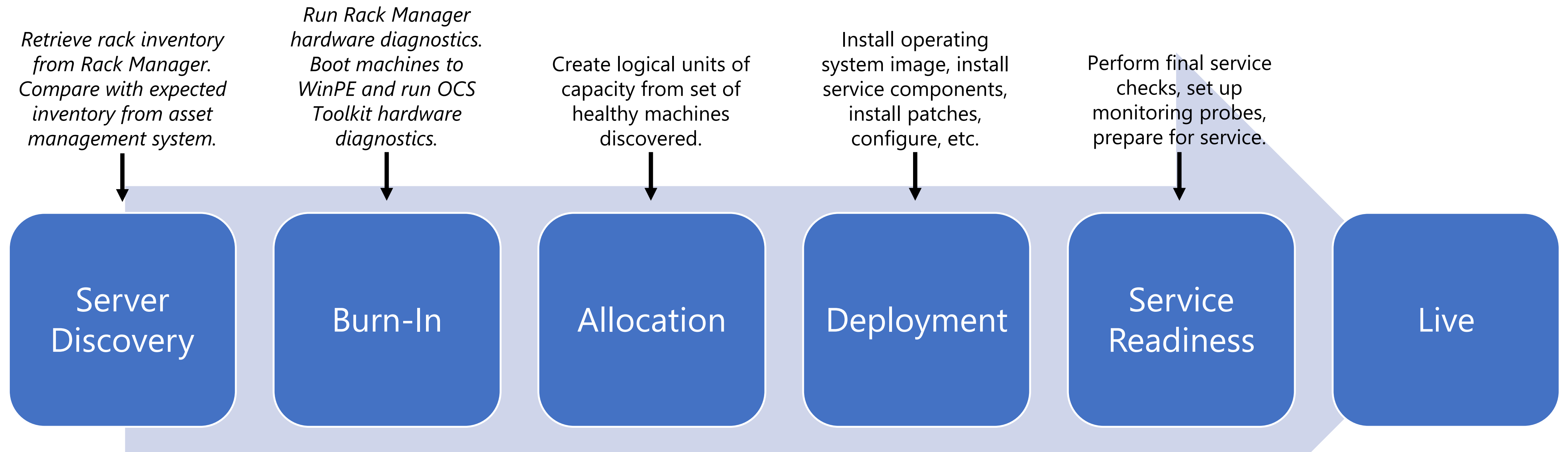lidLBInventoryData InvalidMachineAss oryData InvalidRackInventoryData InvalidRouterInventoryData InvalidVlanInventoryData LbHardDown LbHealthFailur reAaa LbInbandMgtAuthFailureAaaPlus LbInvalidNetvalData LbPortsOnTorAreInvalid LbSshConnectivityFailure LbUplinksFlap dentialExpired MissingVolumeMountpoint Monitoring NATSwitchBack NetworkPortSetting NtfsCorruption OobmBootOrder OobmConnectivity ectivityCredential OobmConnectivityForWcs OobmConnectivityNoOp OobmConnectivityRack OobmController OobmControllerBattery Oobr e OobmCpu OobmDisk OobmEnclosure OobmFan OobmFirmware OobmFPGAFailure OobmHardDiskDisabled OobmHealthControllerCa lthCPU OobmHealthDiskContlrBattery OobmHealthDiskController OobmHealthFan OobmHealthMemory OobmHealthPowerSupply Oobm erSupplyOnline OobmHealthServerOverHeating OobmInventoryData OobmLogFlighting OobmLogin OobmLogMotherboard OobmLogServerOverheating OobmMachineHealthFlighting OobmMemory OobmNetworkSecuritySetting OobmPasswordPolicy OobmPci OobmPowerState OobmPowerSupply OobmPowerSupplyOnline OobmScan OobmScanFailure OobmScanFailureRack OobmServerOverHeating OobmSystemBattery PatchesMissing PatchesMissingOffline PatchesMissingOnline PerformanceHardDisk ProcessRestart RawDisk ReceiveSideScaling RemotingAccessDenied Remoting onnectivity RemotingOutOfMemory RemotingRack RemotingTimeDifference RemotingTrustRelationship RemotingUnknown RnpHostLeafValidation RnpLoadBalancerValidation RnpRouterValidation RoutesNotFound RoutesNotInInventory SerialNumber ServerDiscovery stateConsistency ckInMaintenance TorHardDown TorInbandMgtAuthFailureAaa TorInbandMgtAuthFailureAaaPlus TorInvalidNetvalData TorSsh ailure pedVolume UninitializedDisk UpTime WcsBiosConfig WcsBlade WcsBootOrder WcsDisk WinpeConnectivity WinpeDhcpCheck WireCheckerInventoryData WireCheckerPowerState WireCheckerValidation WireCheckerValidationRack WireCheckerVlan WriteCachingDisabled

Rack / Chassis Manager

OCS Toolkit

Windows Event Log, WMI, etc.

"RepairBox"

Hardware Failure Detection & Repair (cont.)

Test
- **Every 8 hours**
- Run full test suite against every machine in inventory
- Store test results in RepairBox database

Triage
- **Every 1 hour**
- Parse latest test results to identify underlying failures
- Create or update associated failure entries in RepairBox database

Repair
- **Every 4 hours**
- Take machine out of service for repair, if needed
- Perform repair actions (auto-remediation or break/fix ticket)

Repair (cont.)
- **Every 4 hours**
- If initial repair is unsuccessful, perform additional repair actions
- If subsequent repair actions are unsuccessful, flag machine for service engineer attention

Return to Service
- **Every 4 hours**
- Validate repairs and re-configure machine
- Check for any new hardware issues
- Return machine to service

Note: Exchange Online is a highly available service with multiple layers of redundancy that enable us to meet the committed Office 365 SLAs for our customers. These Hardware Failure Detection & Repair timelines are NOT related to the real-time monitoring that happens at the application and service level which, together, ensure uptime of the overall service.
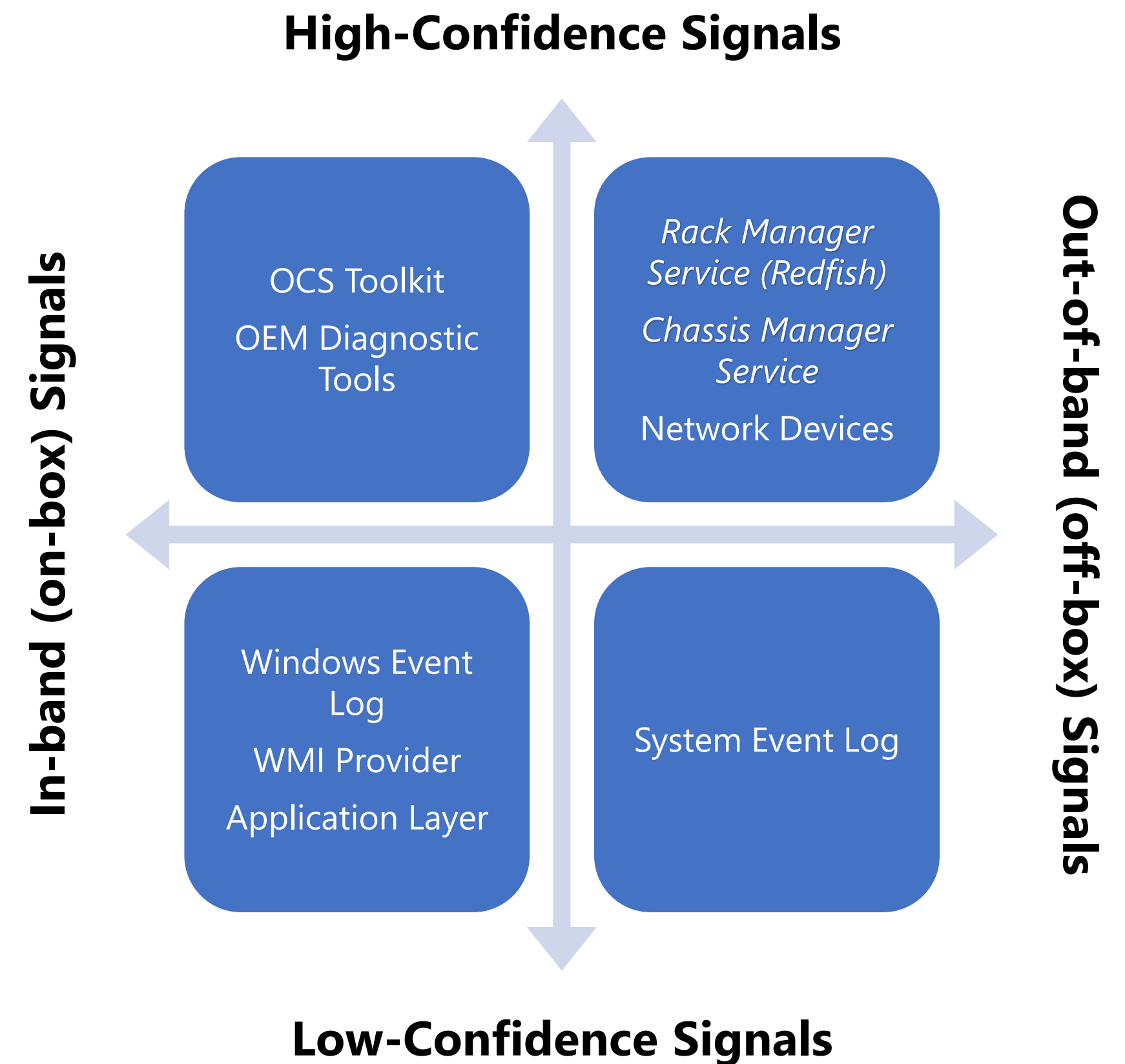
OCP SUMMIT

Open. Together.

Hardware Failure Detection & Repair (cont.)

How to handle low-confidence signals?

Signal aggregation and multi-layered triage (subsystem, machine, rack, etc.).

How to handle "gray" failures?

Prediction model based on observed, historic failures from similar machines.

High-Confidence Signals

In-band (on-box) Signals

Out-of-band (off-box) Signals

OCS Toolkit
OEM Diagnostic Tools

Rack Manager Service (Redfish)
Chassis Manager Service
Network Devices

Windows Event Log
WMI Provider
Application Layer

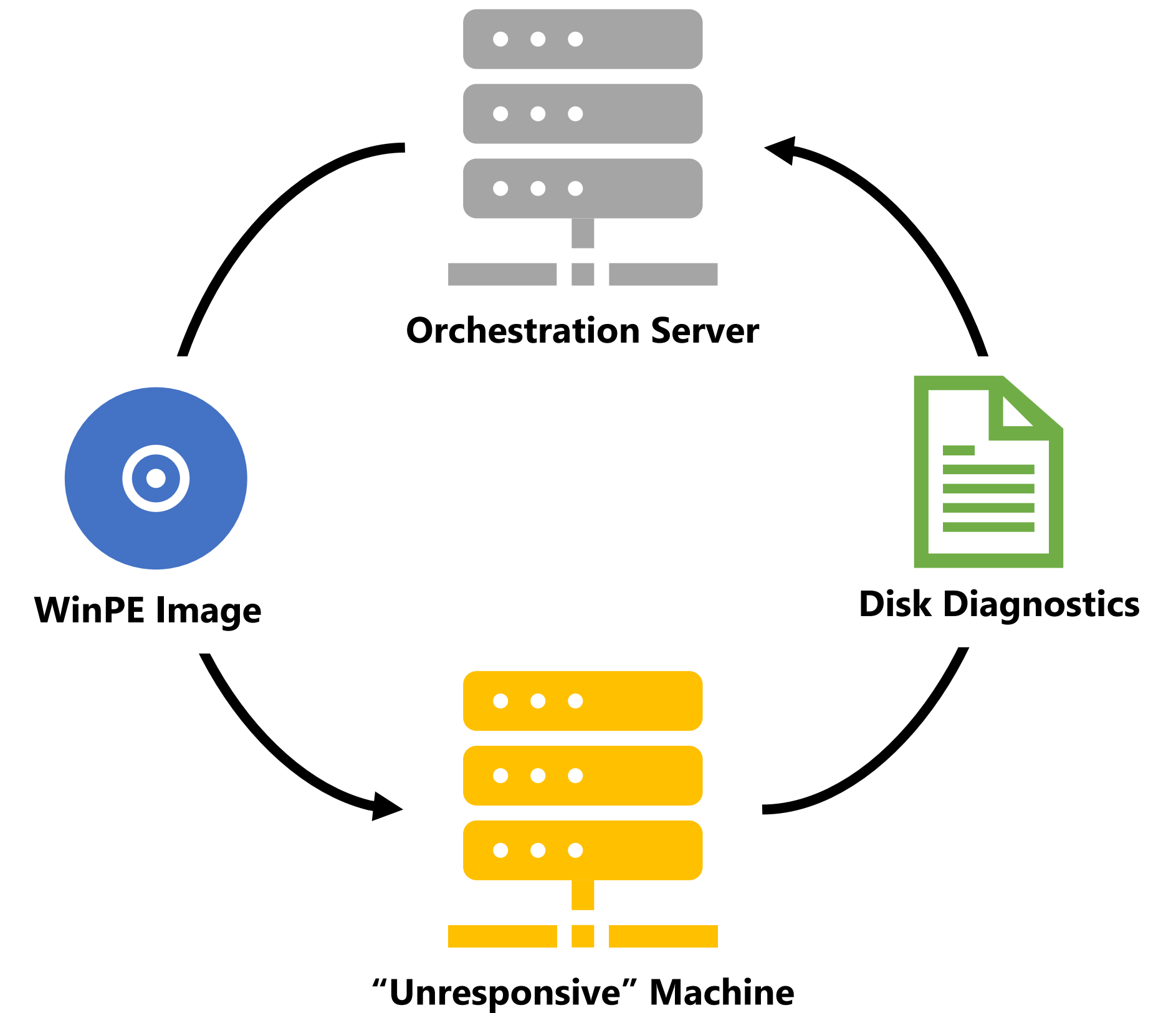System Event Log

Low-Confidence Signals

Disk Failure Detection

No disk failure detection using out-of-band management (Redfish) interface on the WCS platform. Disk failures can only be detected and identified using in-band methods.

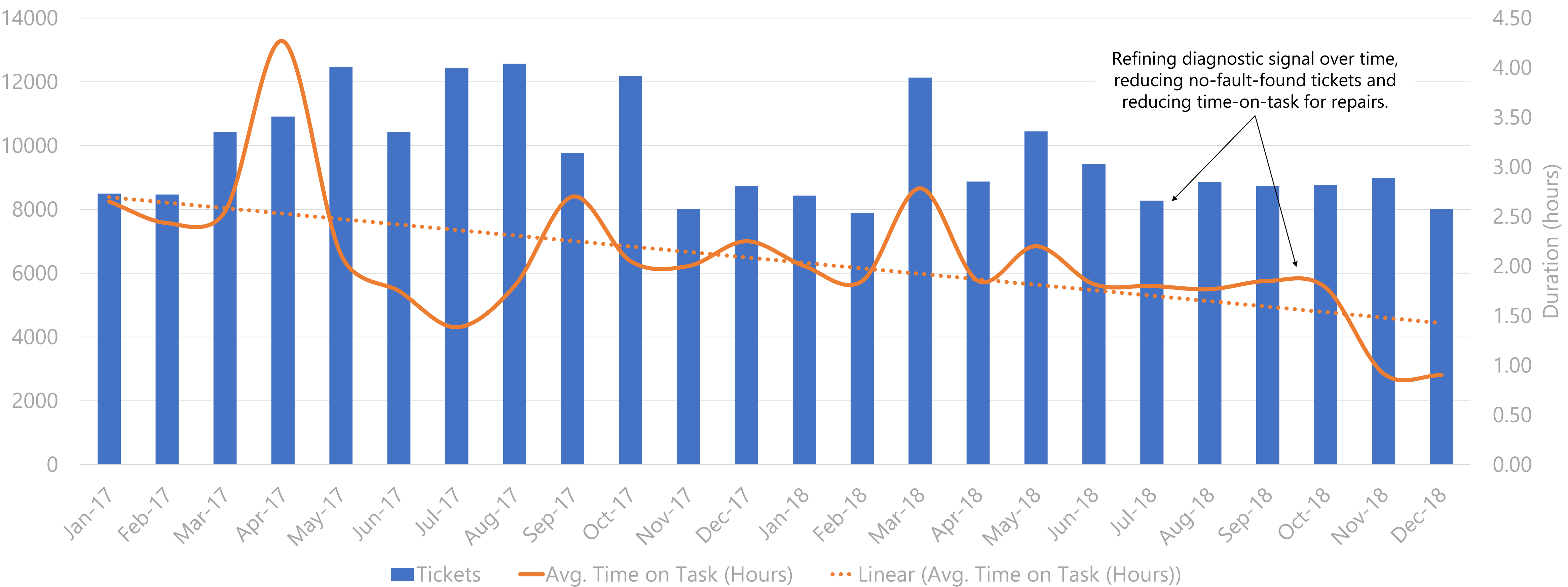**How to handle OS disk failure detection?** Remotely boot "unresponsive" machines to a WinPE diagnostics image and run in-band diagnostic tools.

Orchestration Server

WinPE Image

Disk Diagnostics

"Unresponsive" Machine

OCP SUMMIT

Open. Together.

Break/Fix Tickets

Break/Fix Ticket Volume & Avg. Time on Task

Refining diagnostic signal over time, reducing no-fault-found tickets and reducing time-on-task for repairs.
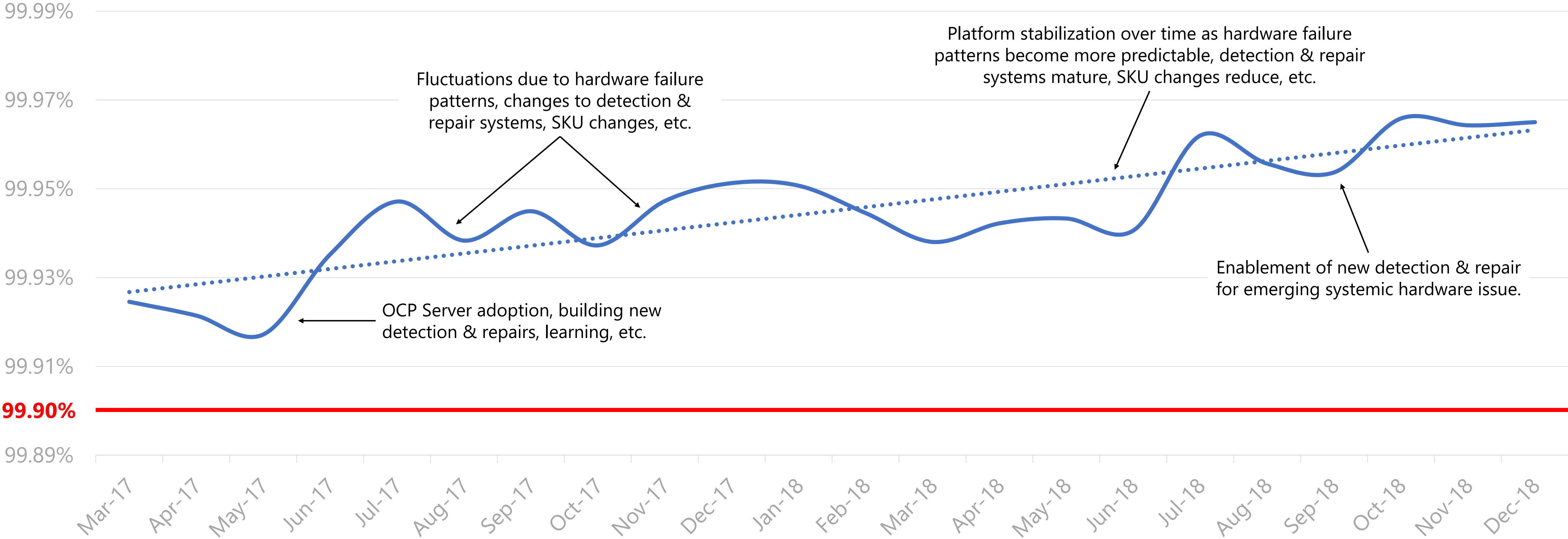
■ Tickets　— Avg. Time on Task (Hours)　⋯⋯ Linear (Avg. Time on Task (Hours))

Note: This data for all servers and SKUs in the Exchange Online fleet, not just the Open Cloud Server platform servers and SKUs.

Open. Together.

Hardware Availability

Monthly Hardware Availability (WCS Gen 5.0 + WCS Gen 6.0)

Fluctuations due to hardware failure patterns, changes to detection & repair systems, SKU changes, etc.

Platform stabilization over time as hardware failure patterns become more predictable, detection & repair systems mature, SKU changes reduce, etc.

OCP Server adoption, building new detection & repairs, learning, etc.

Enablement of new detection & repair for emerging systemic hardware issue.

Note: Exchange Online is a highly available service with multiple layers of redundancy that enable us to meet the committed Office 365 SLAs for our customers. This Hardware Availability metric is NOT a measure of Exchange Online service availability.

Open. Together.

Call To Action

Asks for Hardware Management Community

- **SimpleStorage implementation in Redfish OCP Server profile**

  Expose disk health information in Redfish for OCP servers, so that disk failures can be detected and identified out-of-band. Significant amount of engineering work and infrastructure is needed today to work around this issue, especially for OS disk failures.

- **Hardware Failure Emulation**

  Move away from a "reactive" approach for developing hardware failure detections and towards a "proactive" method in which we can force Redfish APIs and System Event Log (SEL) to emit the associated failure signals for certain failure modes that cannot otherwise be simulated in a lab environment.

- **Standardize OEM entries in System Event Log (SEL)**

  Currently no standardization for how custom SEL entries are written by OEMs and how they are decoded using IPMI. Significant amount of time is invested today in parsing and understanding SEL entries for each OEM and each diagnostic tool, in order to detect hardware failures that are not surfaced by Redfish APIs.

OCP SUMMIT

Open. Together.

Get Involved

Hardware Management working group:
https://www.opencompute.org/wiki/Hardware_Management

Open RMC working group:
https://www.opencompute.org/wiki/Hardware_Management/Open_RMC

MANAGEMENT

Questions?

Matt McCormick

mattmc@microsoft.com

Open. Together.

OCP Global Summit | March 14–15, 2019

OCP SUMMIT

Abstract

Microsoft's Exchange Online service has converged on the Open Cloud Server (OCS) and Project Olympus platforms for its server infrastructure. We are operating a fleet of 100s of 1,000s of OCP servers, with varying generations and SKUs of hardware, in over 100 datacenters worldwide. This session will cover how we discover, provision, monitor, diagnose, and repair our OCP servers, how we have engineered solutions to current manageability limitations, and discuss opportunities for the Hardware Management group to improve this experience.

OCP SUMMIT

Open. Together.