

OPEN

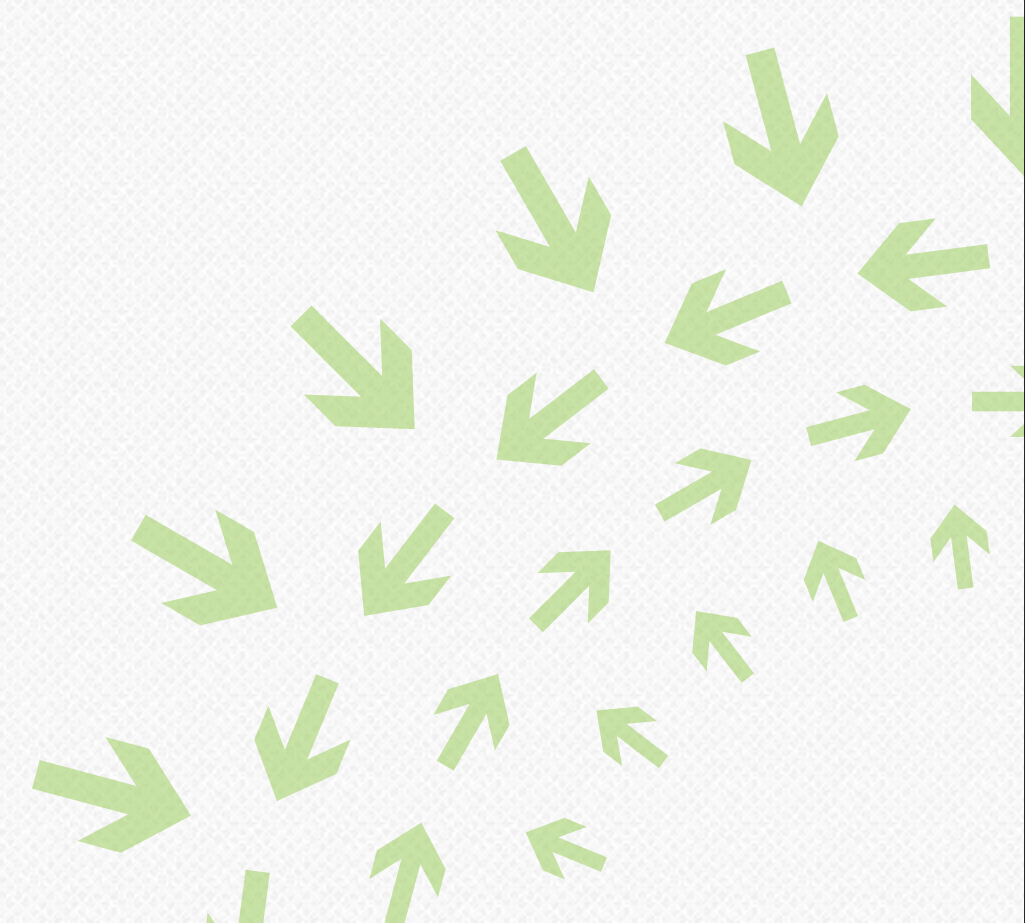
Compute Summit
January 28–29, 2014
San Jose





Type A IPM Controller

Hank Bruning
JBlade
hank@jblade.com



The Goal #1

- Increase the raw data describing the hardware
 - Inventory data on DDR3 /DDR4 and optical
 - More Sensors
 - SFP+ optical module 22 sensors (5 are temperature)
 - CXP optical module 62 sensors (13 are temperature)
 - DDR3 /DDR4 has a single temperature sensor
 - More sensors allow raising DC temperature and get feedback how close to failure you are



Goal #2

- Provide a uniform optical(QSFP+ /CXP), memory DDR3 /DDR4, and chassis identification for:
 - Servers
 - Storage
 - Network switches
- No operating system is required
- Independent of CPU type



Type A IPM Controller Overview

- An implementation of IPMI FRU data and commands to:

- OCP Chassis identification including 1/2 U height
- IPMI Commands that are mandatory
- DDR3/DDR4 inventory and temperature sensors
- Optical XFP/SFP/QSFP/CXP Inventory and control
- IPMI Command to reset IPMI back to system defaults



Type A is not mandatory

- Building block

- A future Type B may do the same thing at a different price point. DMTF ?

- Different Type Deployment scenarios

- First 1000 servers from a new Generation
- Only present in every 10th rack
- A mix of Type A and others in a multi node

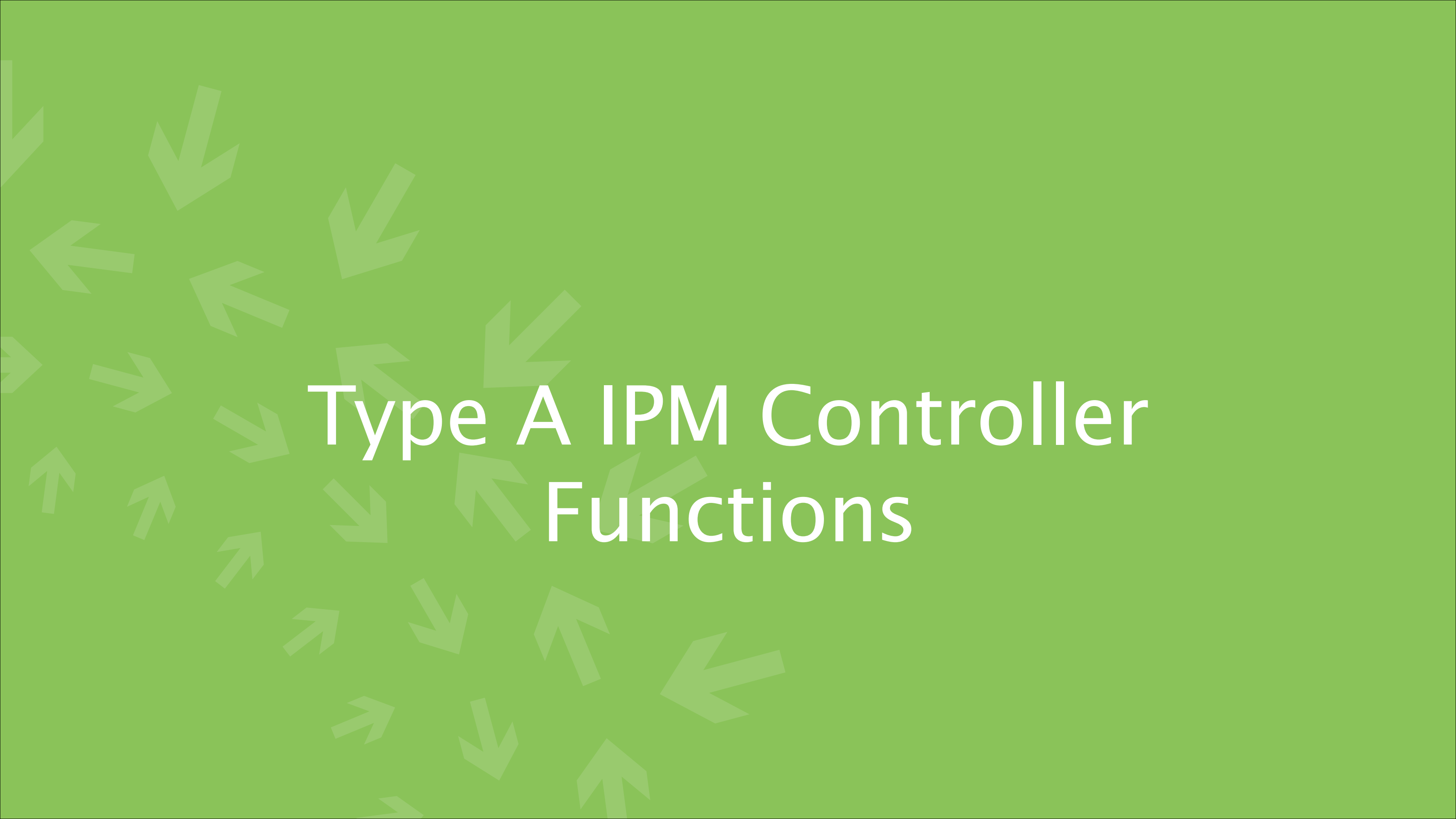


Example of IPM Controller Spec Family

Type	Description
A	IPMI DDR memory/ Optical
B	Same as Type A but DMTF and no IPMI
C	Multi node IPMB Bus Security
D	IPv6 Multicast of sensor readings
E	IPMI PCIe SSD Encryption Key Management

- Multiple specifications allow the Data Center to comparison shop





Type A IPM Controller Functions



Chassis Identification

Problem: Can not identify OCP chassis

- A rack layout can not be drawn
- OCP allows 1/2 U high chassis
 - Minimum size is 1U
- Can not identify a OCP compliant chassis
 - Don't know if OCP IPMI commands should be sent to the chassis



OCP Chassis Identification

- Need to identify Server/Storage/Switch using OCP 1/2 U size.
- Does not allow different heights from and rear. No stair step servers.
- Implemented with FRU Info Multi record





Mandatory IPMI Implementation

Problem. Inconsistent IPMI

- Cost pressure increases the number of incompatible IPMI implementations
- Data Center needs some assurance that new hardware will work with old System Manager



Mandatory IPMI Commands

- Get Channel Authentication Capabilities
 - valid inside and outside RMCP session
- SDRs for Fan speed must contain a max RPM
 - some vendors set it max speed to zero. Can not inform user how close to max speed the fan is running
 - Operator has no data to understand how close to a temperature alarm they are





DDR3 / DDR4 Support

Problem: DDR3 /DDR4 heat

- DDR3 memory varies widely
 - Module case temperature options
 - 85° C with refresh rate 65ms
 - 95° C with refresh rate 32ms
 - Memory speed
 - Is 100% of memory access from single Rank
 - Does module have a heat sink ?
- DC has no view into memory temperatures



DDR3 / DDR4 Slot Inventory

- Server/Storage/Switch has static FRU Info with the quantity and type of memory slots

- Mapping of allows finding total slot count, never changes from IPMI view
- Independent of memory module population
- Inventory includes vendor dependent slot name. Error messages specific to slot.
- Implementation is changed. Now use FRU Device Locator Record. No custom software.



DDR3 /DDR4 Module Inventory

- Map the Serial Presence Detect data to FRU ID
 - Vendor, model, serial number
 - Capacity
 - Speed
 - Temperature rating 85°C or 95°C rating



DDR3 /DDR4 Module Inventory Benefits

- Real time database on memory population.

Allows finding servers:

- with enough capacity to load an O/S and application
- with empty memory slots to upgrade
- with DDR3 /DDR4 modules that can be replaced with higher capacity
- when decommissioning what memory can be removed and reused in new hardware



DDR3 / DDR4 sensors

- IPMI Present/Absent sensor for memory module
- Temperature sensor
 - Alarm thresholds built from 85°/95° C
 - DC can decrease workload on servers with temperature alarms
 - DC more likely to raise air temperature if they know memory modules are within operating bounds





Optical XFP / SFP+ / QSFP+ / CXP Support

Problem: Optical modules vary

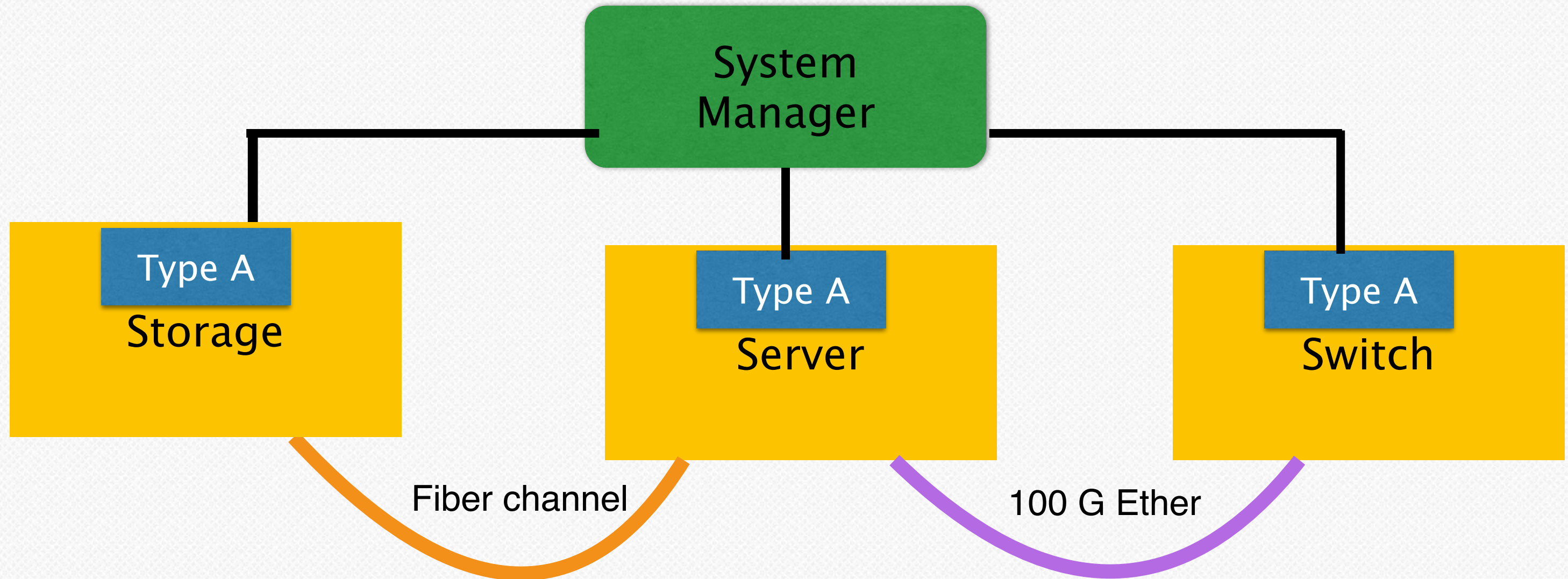
- Protocol independent way to diagnose problems

- Fiber Chanel
- Infiniband
- 100G Ether
- 128G SAS



Leverage IPMI to monitor optical links

- System Manager uses identical diagnostics



Problem: Optical modules vary

- Optical Tx/Rx power varies by temperature and voltage
- Raise room temperature optical links may fail.
- SFF-8636 allows real time measurement of optical link margins



Optical Bay Inventory

- Bays are identified by:
 - FRU ID
 - Location Front/Rear
 - Type XFP/SFP+ /QSFP+ /CXP
 - Row/Column. For drawing highlights
 - Vendor dependent bay name so System Manager provides bay dependent error messages



Optical Module Inventory

- Map 128 Bytes of SFF defined data to FRU ID
 - Vendor ID, Model, Serial number
 - Optical transmit power 1.0, 1.5, 6.5 watts



Optical Module Sensors

- IPMI Present/Absent sensor for optical module
- Detect and process within 20 seconds
- Each optical lane has 5 sensors, two thresholds
 - Transmit optical power
 - Temperature
 - Receive power(two way to measure)
 - Transmit Bias
 - Input voltage





IPMI Reset Command

Problem: No way to reset IPMI variables

- Only current way to do this is reflash BMC
- No quick way to return Server/Storage/Switch to factory default states
- Reflashing slows down testing and deployment
- Data Centers supplying colocated bare metal servers may have customers with IPMI access which needs to be reset



IPMI Reset Command

- Restore IPMI subsystem to defaults
- Limited to IPMI. Not change to BIOS or PMBus
 - Defaults are what ever vendor defines
 - Not happy with the way it's written
 - No OCP vendor independent testing possible
 - No consistent RMCP account, VLAN, IP





Future Changes

Potential Changes to specification

- Move optical to a separate specification
- Add options to define what happens with Event Log Full. Discard new events vs. discard oldest events
- Make RMCP session activation a sensor that gets logged.
 - Discover RMCP account attacks
 - Win for colocation Data Centers supplying bare metal
 - Good for multi-node ?





Questions ?

Ask on the OCP HW Management reflector