Finding the Needle in the Haystack!

Predicting Storage Device Failures in Large-Scale Data Centers

IBM

Fanjing Meng, Hua Ye, Robert Barron, Hongxin Hou, Zeming Zhao, Xiaotian Xu, David Cesarano

David Cesarano
Account Technical Leader for Microsoft
cesarano@ibm.com

Fanjing Meng (Meg)
STSM & CTO
IBM China Systems Lab
mengfj@cn.ibm.com

Hua Ye
Technical Solution Architect
IBM Technology Sales
yehua@cn.ibm.com

Robert Barron
SRE Architect
IBM CIO Hybrid Cloud Platform
brobert@il.ibm.com
@flyingbarron Medium

Hongxin Hou

IBM China Systems Lab

Zeming Zhao

IBM China Development Lab

Xiaotian Xu

IBM China Development Lab

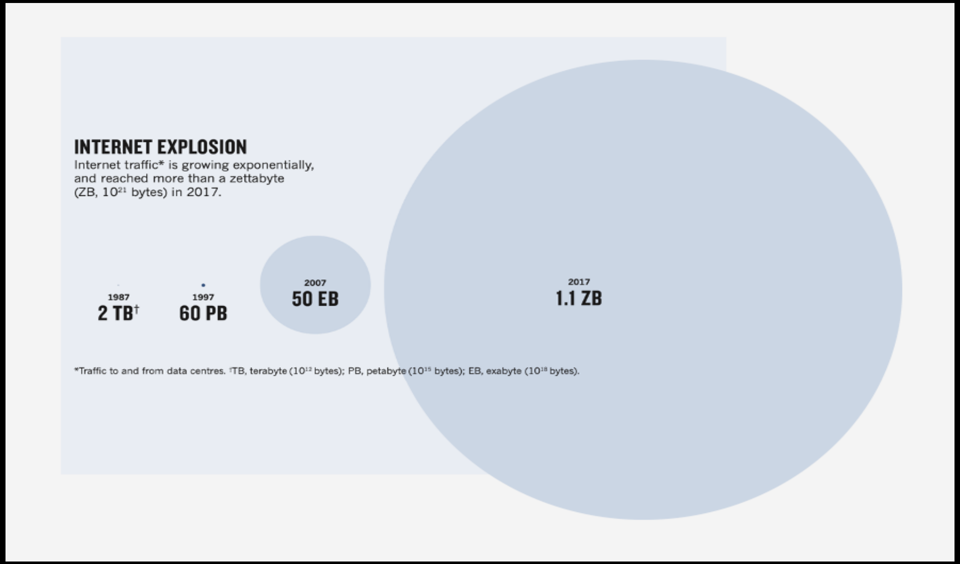
Agenda

- Motivation
- Technical Challenges
- Solution Overview
- Solution Details
- Available Solutions
- Take-aways

Motivation – Data is vital asset and at exponential growth rate

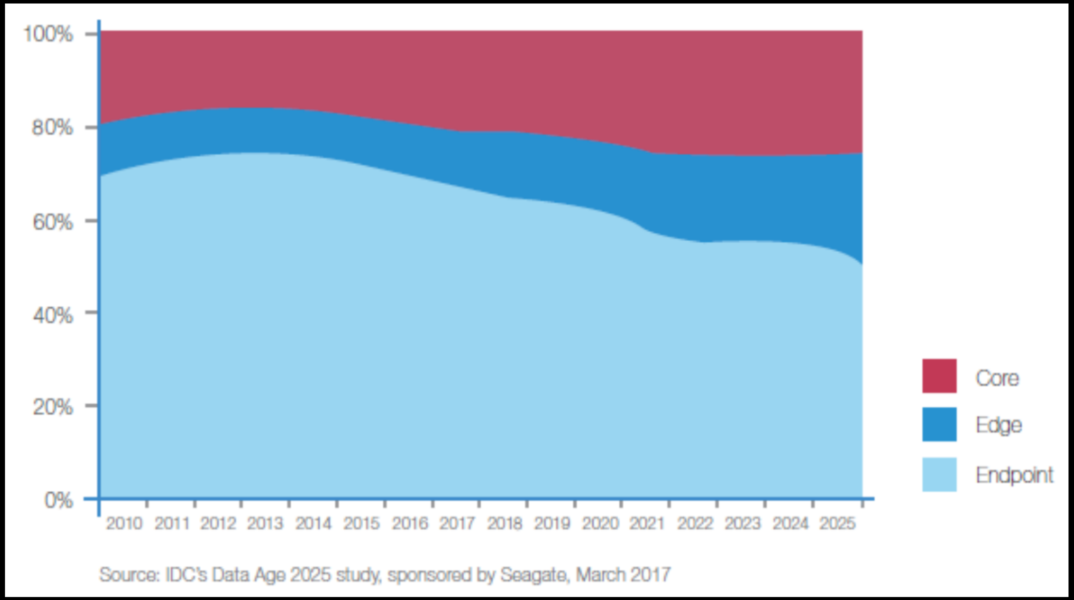
- Data is one of the most important assets in the information era. Data is generated from all the compute devices including sensors, mobiles, cloud/environments at edges, services within data centers etc. Data are collected from the source to the destinations where we consume, manage, store and archive it. It requires IT resources including compute, storage and bandwidth.

Data growing is at an exponential rate



Source: How to stop data centres from gobbling up the world's electricity
https://www.nature.com/articles/d41586-018-06610-y

Data scale at Edge is growing faster and will exceed that at Core



Source: IDC Data Age 2025, http://wwww.dataage2025.com/

Motivation – Failures of storage devices are pervasive

Three Year Comparison of Annual Backblaze Hard Drive Failure Rates

Reporting periods: 1/1/2020 – 12/31/2020, 1/1/2021 – 12/31/2021, 1/1/2022 – 12/31/2022

			2020		2021		2022	
MFG	Model	Size	Drive Count	AFR	Drive Count	AFR	Drive Count	AFR
HGST	HMS5C4040ALE640	4TB	3,100	0.27%	3,429	0.58%	3,723	0.63%
HGST	HMS5C4040BLE640	4TB	12,744	0.27%	12,703	0.31%	12,730	0.41%
HGST	HUH728080ALE600	8TB	1,075	0.29%	1,124	0.64%	1,117	1.43%
HGST	HUH728080ALE604	8TB					94	5.27%
HGST	HUH721212ALE600	12TB	2,600	0.31%	2,600	0.27%	2,606	0.27%
HGST	HUH721212ALE604	12TB	2,506	1.19%	13,138	0.29%	13,165	0.56%
HGST	HUH721212ALN604	12TB	10,830	0.46%	10,818	0.48%	10,769	0.74%
Seagate	ST4000DM000	4TB	18,939	1.41%	18,611	1.80%	18,246	3.45%
Seagate	ST6000DX000	6TB	886	0.23%	886	0.11%	886	0.68%
Seagate	ST8000DM002	8TB	9,772	0.93%	9,718	1.46%	9,523	1.97%
Seagate	ST8000NM000A	8TB					79	0.00%
Seagate	ST8000NM0055	8TB	14,406	1.22%	14,334	1.49%	14,417	2.42%
Seagate	ST10000NM0086	10TB	1,201	1.33%	1,192	2.26%	1,174	3.73%
Seagate	ST12000NM0007	12TB	23,036	1.04%	1,324	2.01%	1,262	4.75%
Seagate	ST12000NM0008	12TB	19,287	1.01%	20,201	1.08%	19,821	2.02%
Seagate	ST12000NM001G	12TB	7,130	0.84%	12,171	0.52%	12,623	0.94%
Seagate	ST14000NM001G	14TB	5,987	1.04%	10,738	1.03%	10,751	1.18%
Seagate	ST14000NM0138	14TB	360	0.00%	1,611	4.79%	1,519	5.70%
Seagate	ST16000NM001G	16TB	59	1.71%	10,861	1.11%	20,393	0.86%
Seagate	ST16000NM002J	16TB					310	1.44%
Toshiba	MD04ABA400V	4TB	99	2.01%	97	2.04%	94	3.13%
Toshiba	MG07ACA14TA	14TB	21,046	0.91%	38,214	0.77%	38,182	1.01%
Toshiba	MG07ACA14TEY	14TB	160	0.00%	462	1.66%	552	1.58%
Toshiba	MG08ACA16TA	16TB					3,751	0.58%
Toshiba	MG08ACA16TE	16TB			5,985	0.91%	5,936	1.57%
Toshiba	MG08ACA16TEY	16TB	1,014	0.00%	2,367	0.70%	5,286	0.64%
WDC	WUH721414ALE6L4	14TB	6,002	0.16%	8,408	0.43%	8,410	0.12%
WDC	WUH721816ALE6L0	16TB			1,767	0.14%	2,701	0.12%
WDC	WUH721816ALE6L4	16TB					10,801	0.36%
		Totals:	162,239	0.93%	202,759	1.01%	230,921	1.37%

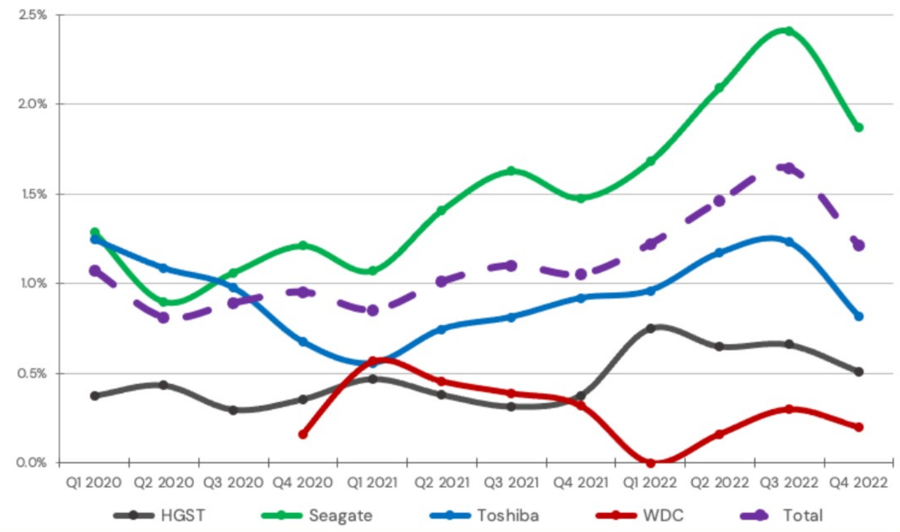
Backblaze

https://www.backblaze.com/blog/backblaze-drive-stats-for-2022/

For the 2022 Disk Stat report, Backblaze tracked 230,921 hard drives grouped into 20 different models to analyze.

Backblaze Quarterly Hard Drive Annualized Failure Rates by Manufacturer

Annualized failure rates for each quarter are computed based on the data from that quarter

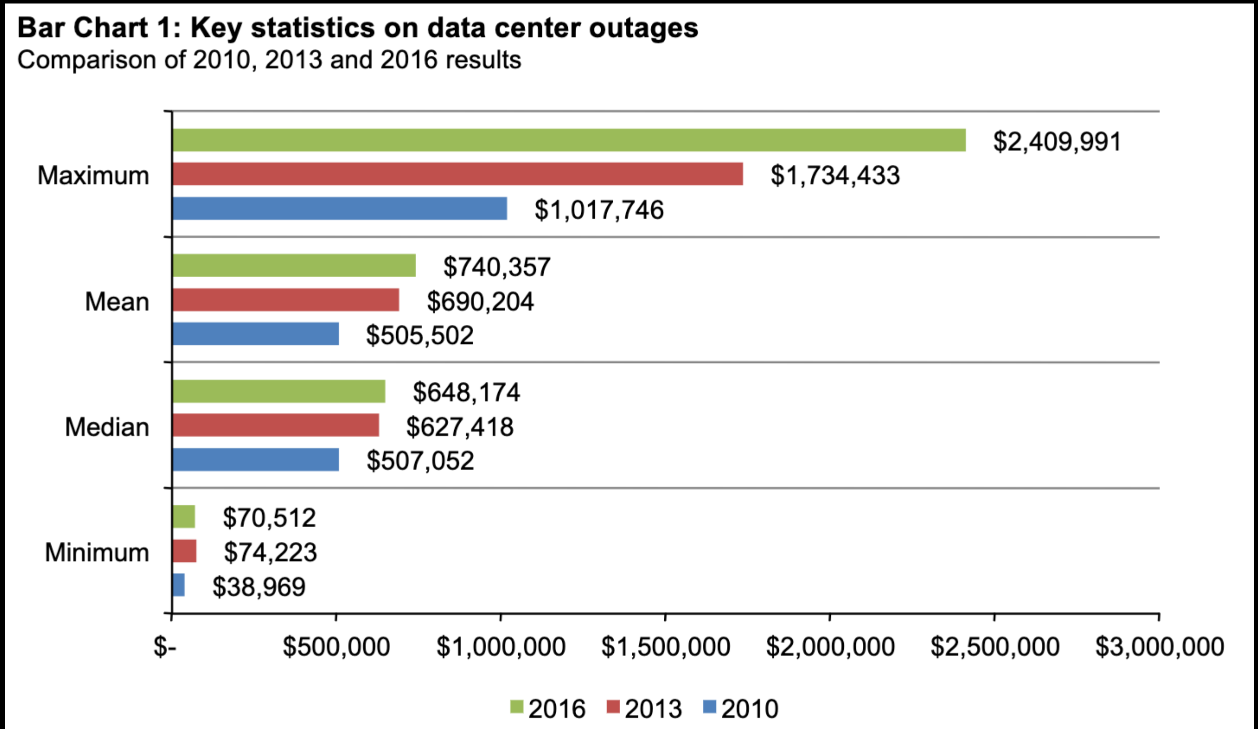


	Q1 2020	Q2 2020	Q3 2020	Q4 2020	Q1 2021	Q2 2021	Q3 2021	Q4 2021	Q1 2022	Q2 2022	Q3 2022	Q4 2022
HGST	0.37%	0.43%	0.30%	0.35%	0.47%	0.38%	0.31%	0.37%	0.75%	0.65%	0.66%	0.51%
Seagate	1.29%	0.90%	1.06%	1.21%	1.07%	1.41%	1.63%	1.48%	1.68%	2.09%	2.41%	1.87%
Toshiba	1.24%	1.09%	0.98%	0.67%	0.56%	0.75%	0.81%	0.92%	0.96%	1.17%	1.23%	0.82%
WDC				0.16%	0.57%	0.46%	0.39%	0.32%	0.00%	0.16%	0.30%	0.20%
Total	**1.07%**	**0.81%**	**0.89%**	**0.95%**	**0.85%**	**1.01%**	**1.10%**	**1.05%**	**1.22%**	**1.46%**	**1.64%**	**1.21%**

Backblaze

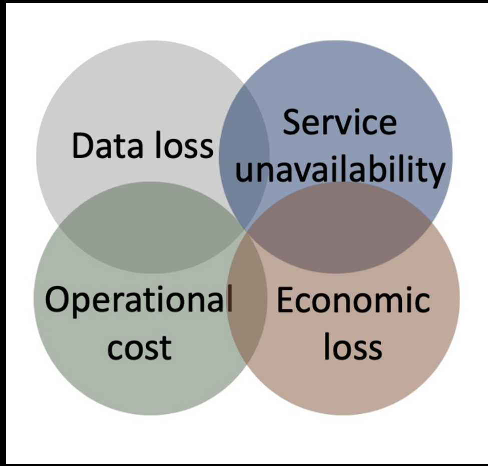
Motivation – Failures cause huge negative impact

The maximum cost has more than doubled over six years from just over $1 million to $2.4 million (a 34 percent increase since the last study).



Source: Cost of Data Center Outages – Ponemon Institute
https://www.vertiv.com/globalassets/documents/reports/2016-cost-of-data-center-outages-11-11_51190_1.pdf

Storage devices failures are costly.

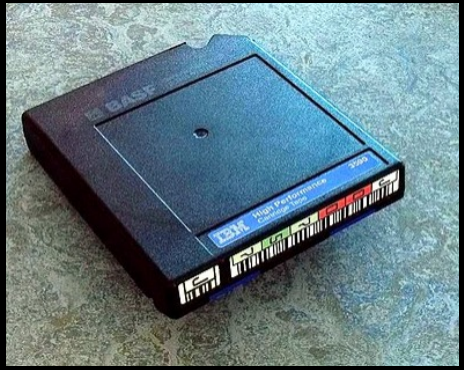


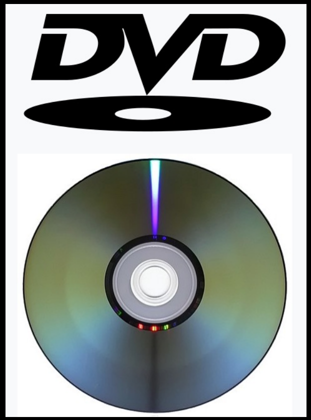
https://medium.com/genaro-network/tencent-was-claimed-ten-million-for-data-loss-due-to-cloud-hard-drive-glitch-344a26449fe2

Technical Challenges - Diversity

Enterprises are using a mix of storage devices/medias to satisfy their performance, management, and archive demands :

- Hard Disk Drives (HDD)
- Solid State Drives (SSD)
- Tape
- Optical Disk Drives (CD,DVD, Blu-Ray DVD)



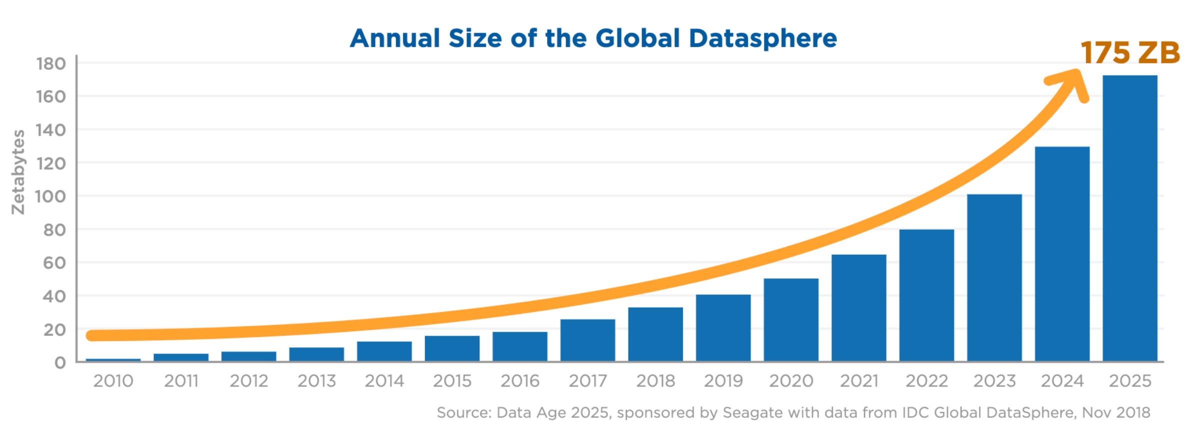


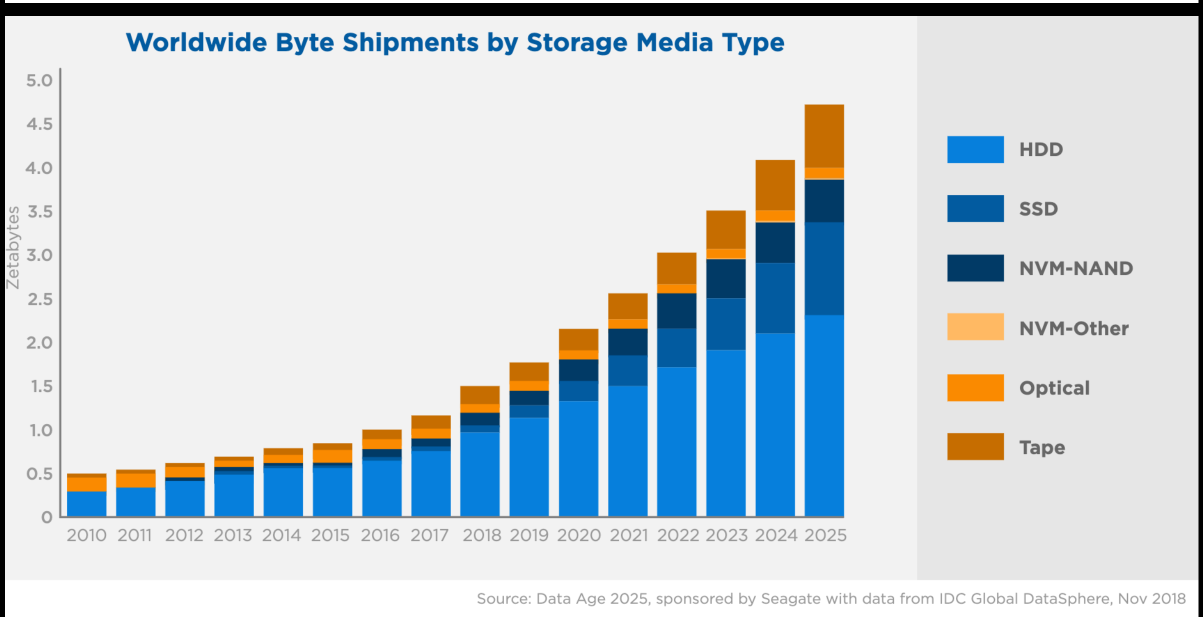
Source: wikipedia

Technical Challenges - Scale

- According to IDC, worldwide data will reach to 175 zettabytes by 2025.

- These large volume of data will storge in a huge amounts of various data storage medias.



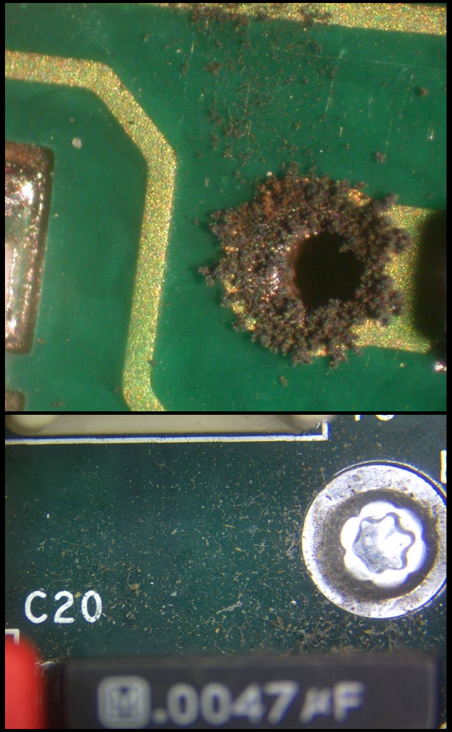


Source: The Digitization of the World - From Edge to Core, https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

Technical Challenges – Complexity

Failures can be caused by many factors:

- Environmental Factors:
  - Temperature, Humidity, Air Quality, Power stability, physical vibration, ...
- Mechanical failures:
  - drive motor or read/write heads failures of HDDs ..
- Media failures:
  - track damage or magnetic surface corrosion of HDDs, flash ...
- Electronic failure:
  - damaged circuit boards or controllers damaged, short-circuited, failures of flash memory chips ...
- Software errors:
  - OS errors, driver conflicts, or malware
- Data corruption:
  - Virus infections, file system errors, data loss
- Aging issues:



Technical Challenges – Multiple Dimensions of Metrics

High-dimensional metrics are monitored in real-time:

- SMART (Self-Monitoring Analysis and Reporting Technology):
  - A monitoring system to detect and report various indicators of hard disk drives (HDDs) and solid-state drives (SSDs).
- Storage performance data:
  - Measurements of how well storage devices perform, especially hard drives. e.g. IOPS, MTBF, MTTR, read/write speed, and etc.
- Environmental monitoring data
  - metrics of environments deployed these disk drives e.g. temperature, humility, air qualities
- Meta data & Maintenance data
  - Vendors, type, maintenance logs

SMART metric examples

ID	Attribute name	Ideal	!
01 0x01	Read Error Rate	Low	
02 0x02	Throughput Performance	High	
03 0x03	Spin-Up Time	Low	
04 0x04	Start/Stop Count		
05 0x05	Reallocated Sectors Count	Low	[25][26][27
06 0x06	Read Channel Margin		
07 0x07	Seek Error Rate	Varies	
08 0x08	Seek Time Performance	High	
09 0x09	Power-On Hours		
10 0x0A	Spin Retry Count	Low	[33]
11 0x0B	Recalibration Retries or Calibration Retry Count	Low	
12 0x0C	Power Cycle Count		
13 0x0D	Soft Read Error Rate	Low	
22 0x16	Current Helium Level	High	
23 0x17	Helium Condition Lower		
24 0x18	Helium Condition Upper		
170 0xAA	Available Reserved Space		
171 0xAB	SSD Program Fail Count		
172 0xAC	SSD Erase Fail Count		
173 0xAD	SSD Wear Leveling Count		
174 0xAE	Unexpected Power Loss Count		
175 0xAF	Power Loss Protection Failure		
176 0xB0	Erase Fail Count		
177 0xB1	Wear Range Delta		
178 0xB2	Used Reserved Block Count		
179 0xB3	Used Reserved Block Count Total		
180 0xB4	Unused Reserved Block Count Total		
181 0xB5	Program Fail Count Total or Non-4K Aligned Access Count	Low	
182 0xB6	Erase Fail Count		
183 0xB7	SATA Downshift Error Count or Runtime Bad Block	Low	
184 0xB8	End-to-End error / IOEDC	Low	[45]
185 0xB9	Head Stability		
186 0xBA	Induced Op-Vibration Detection		
187 0xBB	Reported Uncorrectable Errors	Low	[47]
188 0xBC	Command Timeout	Low	[48]
189 0xBD	High Fly Writes	Low	
190 0xBE	Temperature Difference or Airflow Temperature	Varies	
191 0xBF	G-sense Error Rate	Low	
192 0xC0	Power-off Retract Count, Emergency Retract Cycle Count (Fujitsu),[50] or Unsafe Shutdown Count	Low	
193 0xC1	Load Cycle Count or Load/Unload Cycle Count (Fujitsu)	Low	
194 0xC2	Temperature or Temperature Celsius	Low	
195 0xC3	Hardware ECC Recovered	Varies	
196 0xC4	Reallocation Event Count[50]	Low	[4]
197 0xC5	Current Pending Sector Count[50]	Low	[4][45][48]
198 0xC6	(Offline) Uncorrectable Sector Count[50]	Low	[4][25]
199 0xC7	UltraDMA CRC Error Count	Low	
200 0xC8	Multi-Zone Error Rate [64]	Low	
200 0xC8	Write Error Rate (Fujitsu)	Low	
201 0xC9	Soft Read Error Rate or TA Counter Detected	Low	[66]
202 0xCA	Data Address Mark errors or TA Counter Increased	Low	
203 0xCB	Run Out Cancel	Low	
204 0xCC	Soft ECC Correction	Low	
205 0xCD	Thermal Asperity Rate	Low	
206 0xCE	Flying Height		
207 0xCF	Spin High Current	Low	
208 0xD0	Spin Buzz		
209 0xD1	Offline Seek Performance		
210 0xD2	Vibration During Write		
211 0xD3	Vibration During Write		
212 0xD4	Shock During Write		
220 0xDC	Disk Shift	Low	
221 0xDD	G-Sense Error Rate	Low	
222 0xDE	Loaded Hours		
223 0xDF	Load/Unload Retry Count		
224 0xE0	Load Friction	Low	
225 0xE1	Load/Unload Cycle Count	Low	
226 0xE2	Load 'In'-time		
227 0xE3	Torque Amplification Count	Low	
228 0xE4	Power-Off Retract Cycle	Low	
230 0xE6	GMR Head Amplitude (magnetic HDDs), Drive Life Protection Status (SSDs)		
231 0xE7	Life Left (SSDs) or Temperature		
232 0xE8	Endurance Remaining or Available Reserved Space		
233 0xE9	Media Wearout Indicator (SSDs) or Power-On Hours		
234 0xEA	Average erase count AND Maximum Erase Count		
235 0xEB	Good Block Count AND System(Free) Block Count		
240 0xF0	Head Flying Hours or '*Transfer Error Rate*' (Fujitsu)		
241 0xF1	Total LBAs Written		
242 0xF2	Total LBAs Read		
243 0xF3	Total LBAs Written Expanded		
244 0xF4	Total LBAs Read Expanded		
249 0xF9	NAND Writes (1GiB)		
250 0xFA	Read Error Retry Rate	Low	
251 0xFB	Minimum Spares Remaining		
252 0xFC	Newly Added Bad Flash Block		
254 0xFE	Free Fall Protection	Low	

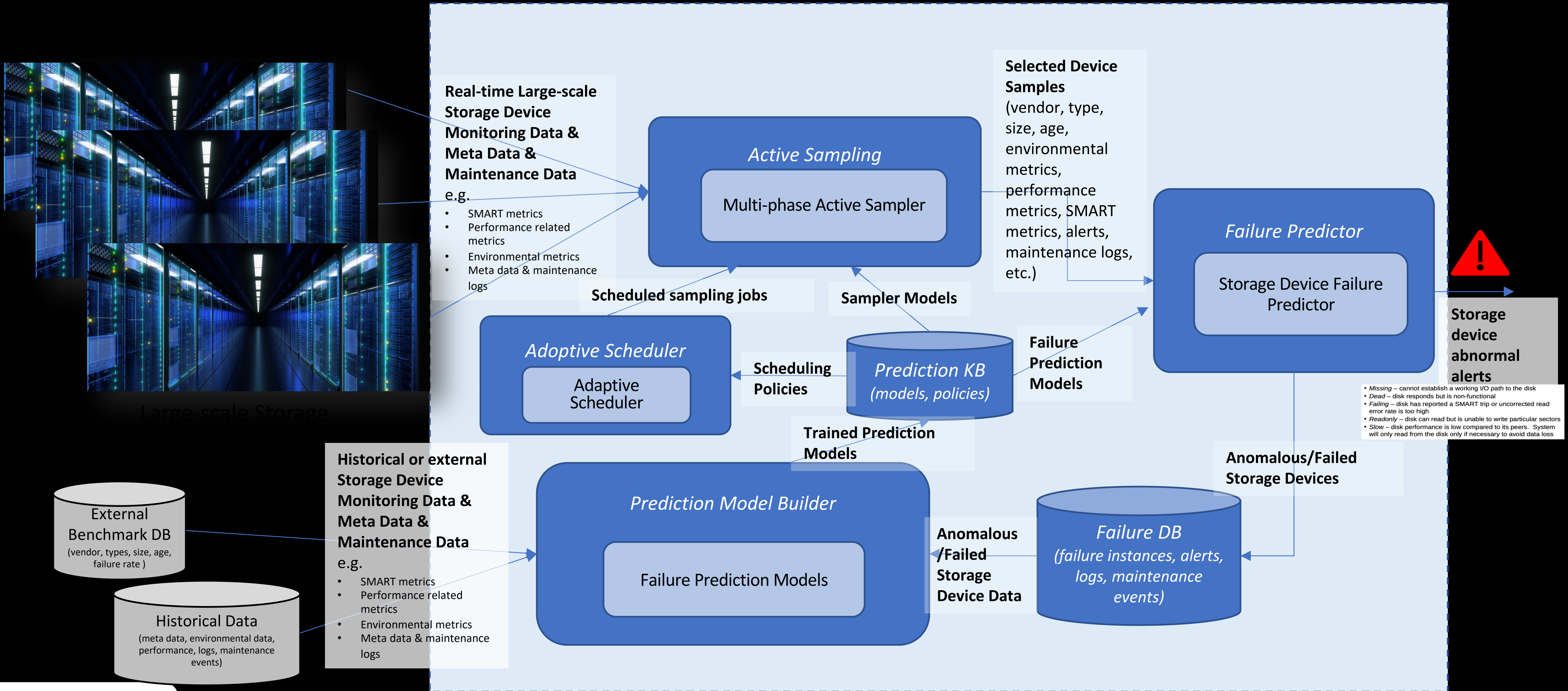
Known Solutions

- Threshold-based or rule-based approaches
  - To set thresholds based on selected metrics, e.g. SMART 05- Reallocated Sectors Count > 258 based on manufacturer recommendations, expert knowledge, or relevant research.
- Statistics-based approaches
  - To uses historical failure data and statistical analysis techniques to predict the failure probability of disk drives based on metrics such as drive operating time, error rates, temperature, etc.
- Learning-based approaches
  - To build machine learning or deep learning models e.g. decision trees, random forests, support vector machines, and neural networks based on historical data to predict anomalies or life span

Problems and Value Propositions

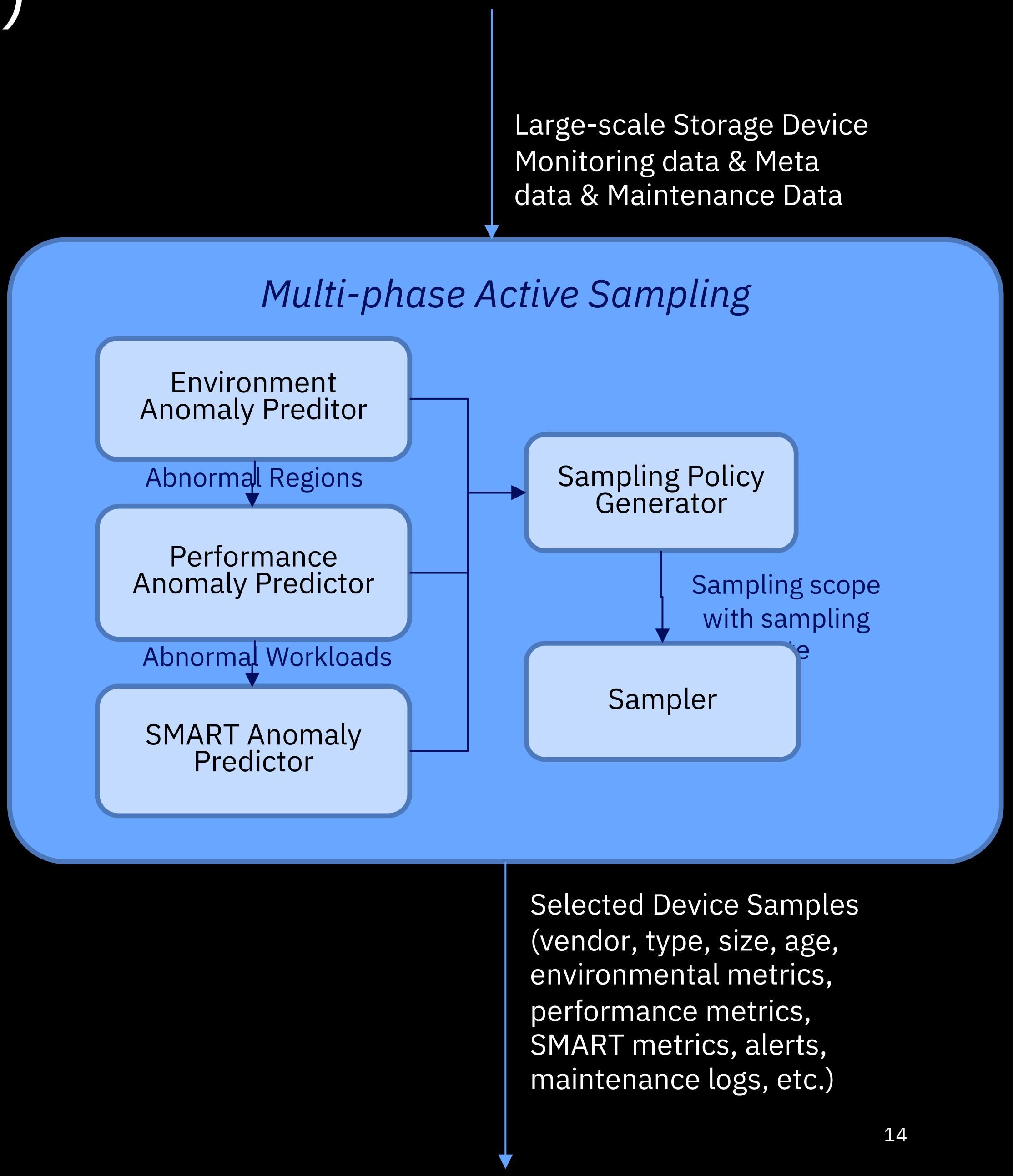
- Problem to Solve:
  - Detect/predict anomalies of massive storage devices are costly and time-consuming.
    - High dimensional monitoring metrics: SMART, Software metrics, Environmental metrics
    - Fast speed of data: data collected at second/minute level
    - Large-scale and diverse monitored objects: HDDs, SDDs, Tapes, etc.
    - Need large real-time detection clusters to detect anomalies and predict failures
- A Novel approach to anomaly detection:
  - **Balance of accuracy, performance, sustainability, and cost** which consider the following aspects:
    - **Active Sampling**: to select the highly representative or highly risky devices for data collection and analysis
    - **Phase-based**: to apply multi-phase filtering approach to reduce scale of data collection & analysis.
    - **Learning-enabled**: apply machine learning/deep learning to continuously improve quality with learning.

Our Solutions – Active Phase-based Sampling Approach



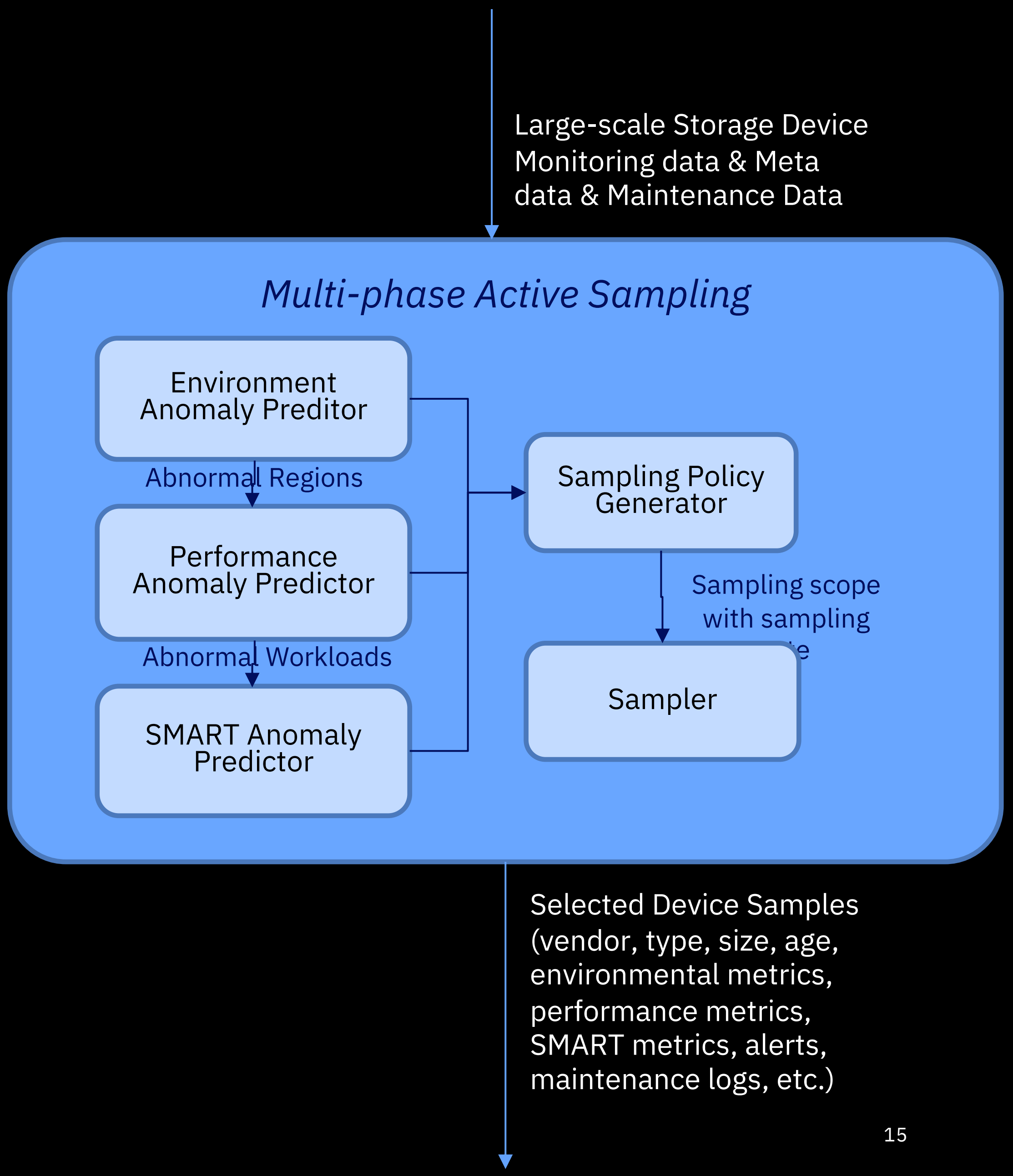
How it works – Active Sampler (1)

- Motivations:
  - To dramatically reduce the data collection and analysis scope and therefore reduce the cost and improve the performances
- Capability:
  - To filter highly representative or highly risky devices for further detailed storage device failure prediction based on failure rate
- Input:
  - SMART metrics
  - Performance related metrics
  - Environmental metrics
  - Meta data & maintenance logs
- Output:
  - Selected samples



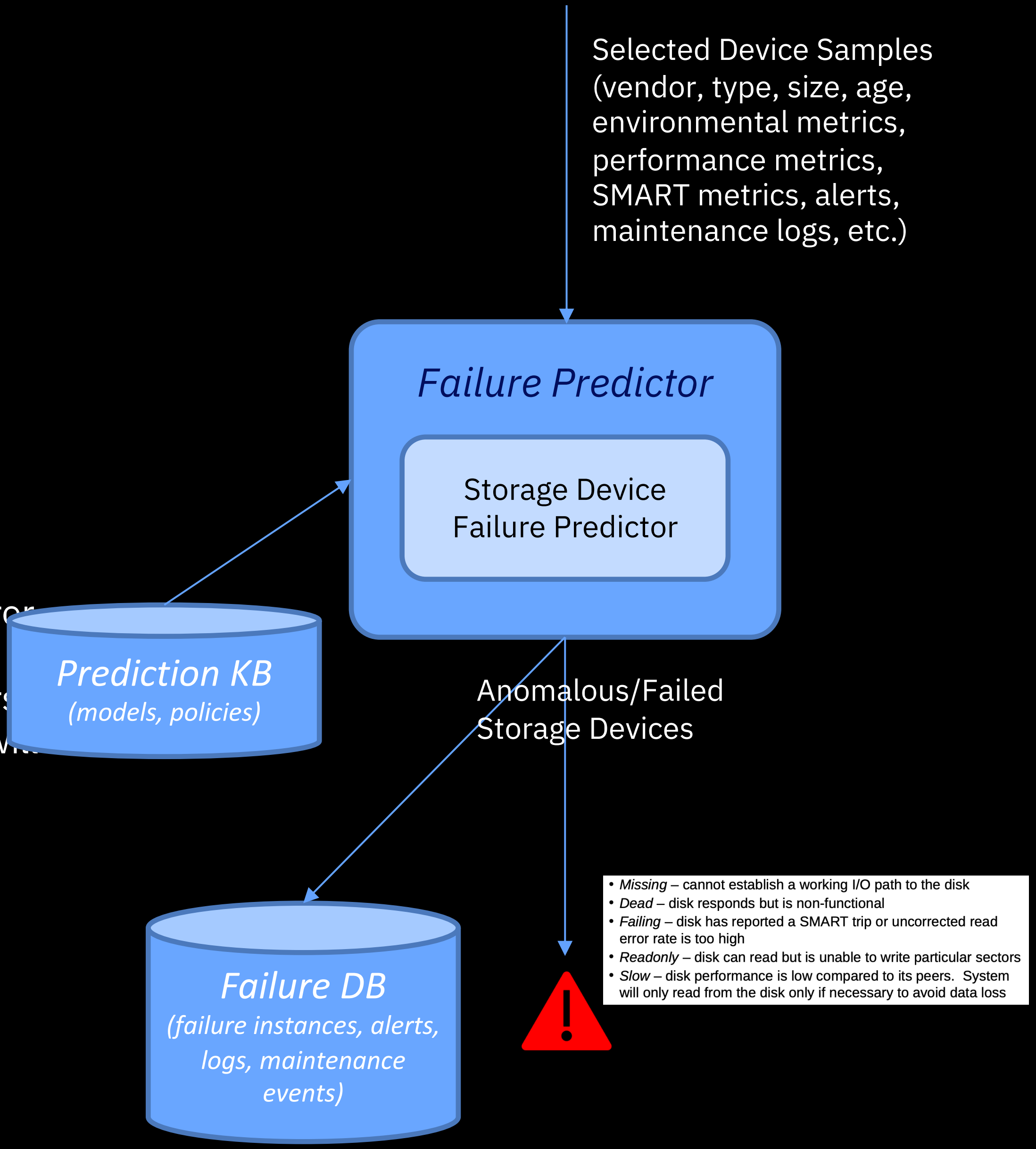
How it works – Active Sampler (2)

- Active sampling mechanism:
  - Anomaly Predictor:
    - Environmental anomaly prediction (classification models + regression models + clustering models + heuristic models)
    - Performance anomaly prediction (classification models + regression models + clustering models + heuristic models)
    - SMART anomaly prediction (classification models + regression models + clustering models + heuristic models)
  - Sampling Policy Generator:
    - Determine sample scope based on prediction results
  - Scope + sampling rate (server room, row, rack, server, storage server, disks, tapes)
  - Sampler
    - Randomly select representative and/or high-risk samples based on instances in failure base



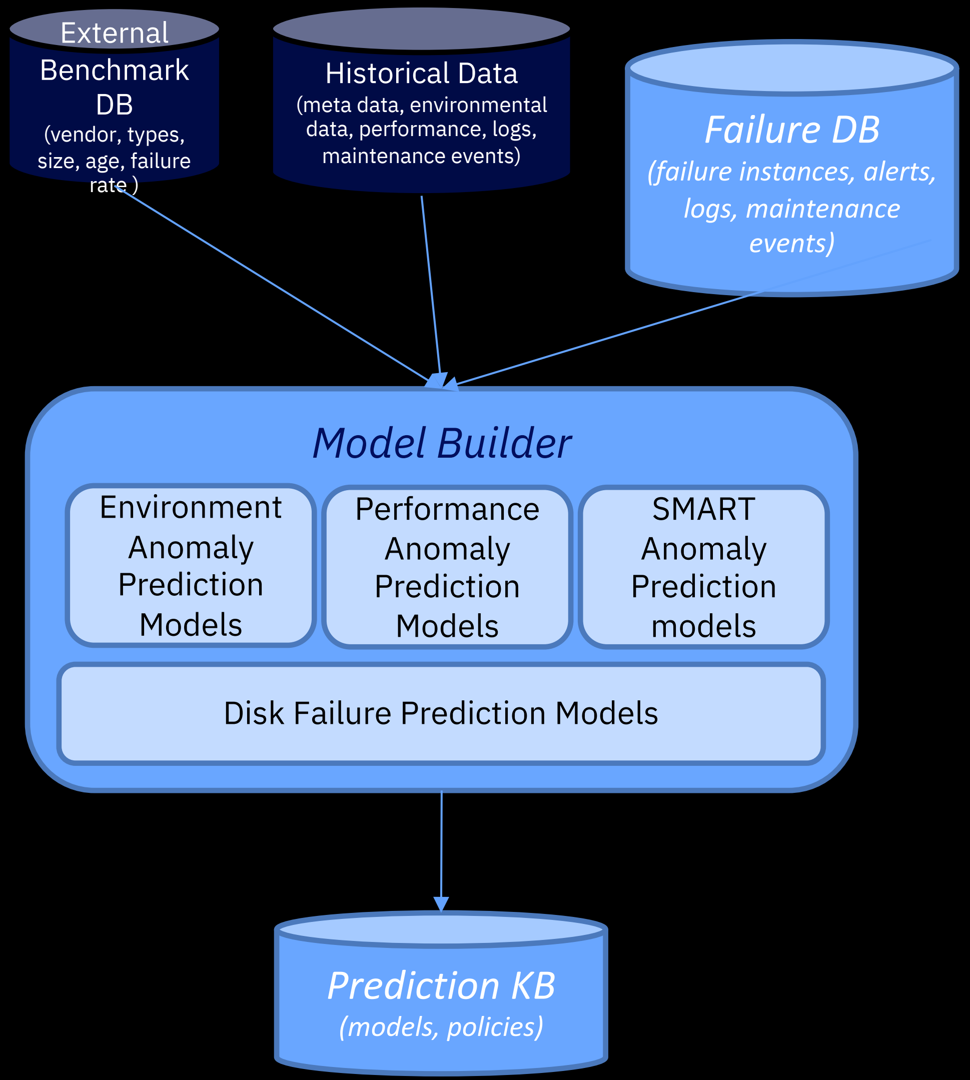
How it works – Failure Prediction

- Capability:
  - To predict storage device failures based on multiple data sources
- Input:
  - SMART metrics
  - Performance related metrics
  - Environmental metrics
  - Meta data & maintenance logs
- Output:
  - Predicted storage failures with failure modes and confidence levels
    - Missing – cannot establish a working I/O path to the disk
    - Dead – disk responds but is non-functional
    - Failing – disk has reported a SMART trip or uncorrected read error rate is too high
    - Readonly – disk can read but is unable to write particular sectors
    - Slow – disk performance is low compared to its peers. System will only read from the disk only if necessary to avoid data loss
- Predict Device Failures of selected samples
  - Apply AI/ML models to predict device failures based on multi-dimensional data
    - Environmental, performance, SMART, meta data, ops and maintenance data



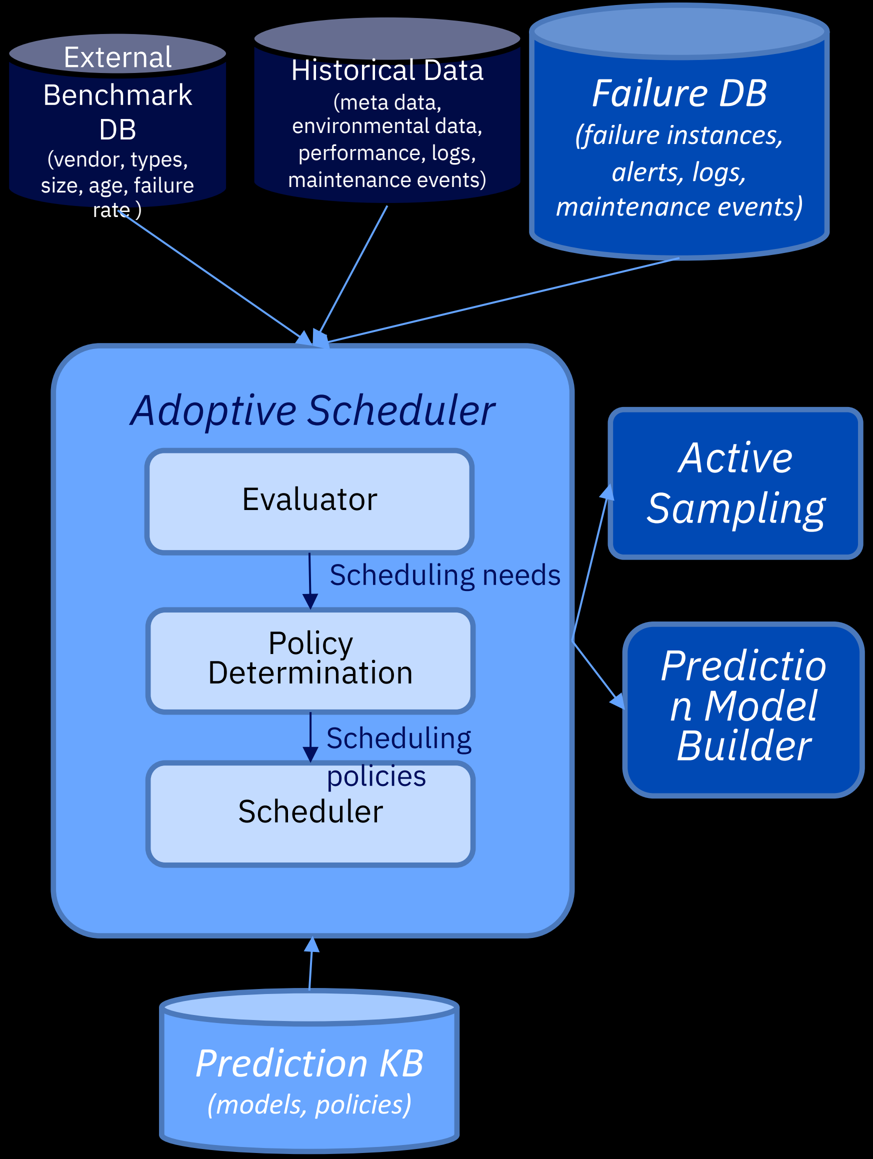
How it works – Benchmark Model Builder

- Capability:
  - To build models for disk failure predictions
- Input:
  - Historical data
  - Benchmark Data
  - Failure DB
- Output:
  - Prediction models
- Detailed approach:
  - Environmental data
    - Environmental Anomaly Prediction models
  - Performance data
    - Performance Anomaly Prediction models
  - SMART data
    - SMART Anomaly Prediction models
  - All Data
    - Storage Device Failure Prediction models



How it works – Adaptive Scheduler (1)

- Capability:
    - To schedule the sampling and model training on demand when new training data available or model drifting
- Input:
    - Historical Data
    - Benchmark Data
    - Failure DB
    - Scheduling Policies
- Output:
    - Scheduled active sampling and model training

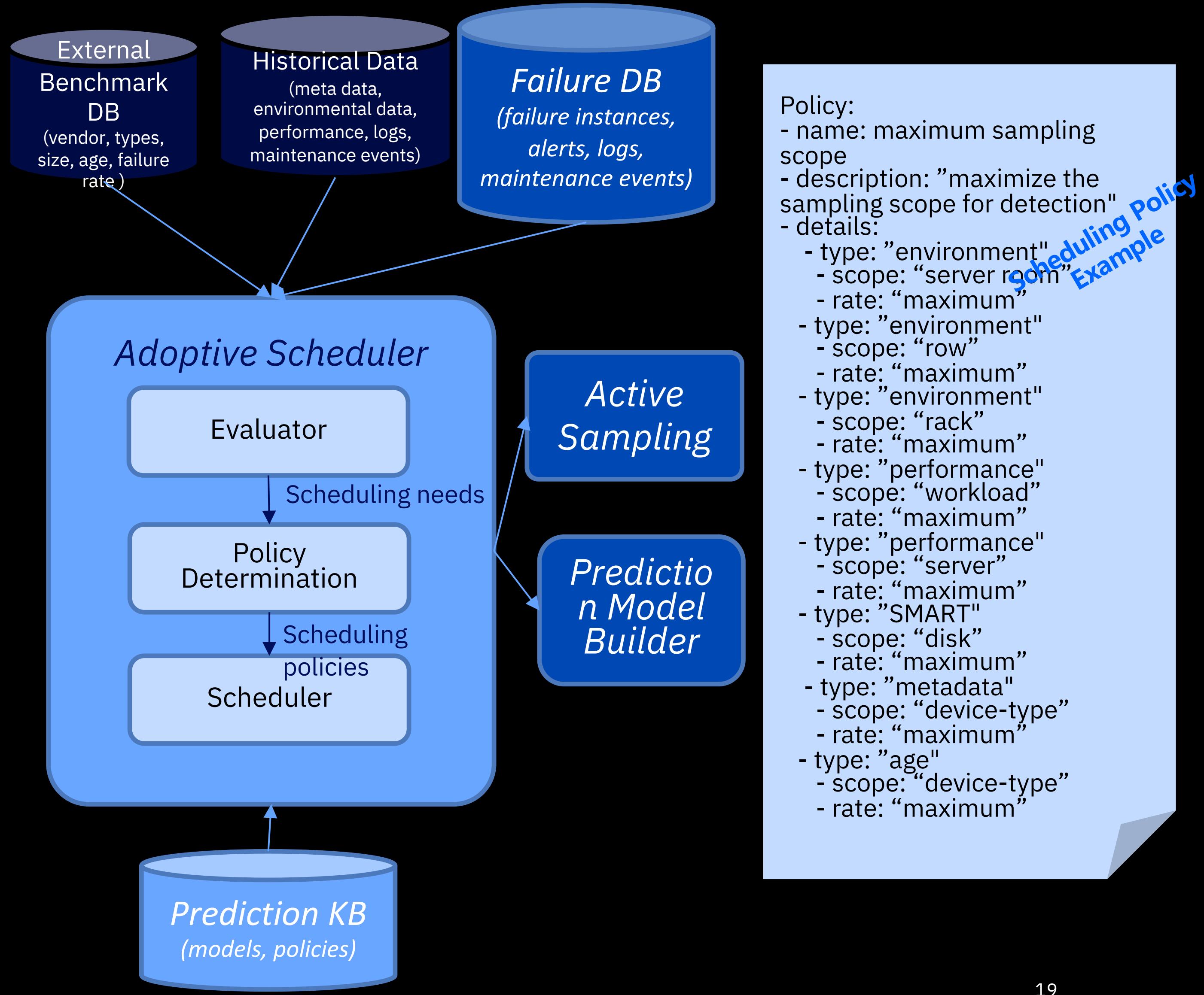


Scheduling Policy Example

Policy:
- name: maximum sampling scope
- description: ”maximize the sampling scope for detection"
- details:
    - type: ”environment"
        - scope: “server room”
        - rate: “maximum”
    - type: ”environment"
        - scope: “row”
        - rate: “maximum”
    - type: ”environment"
        - scope: “rack”
        - rate: “maximum”
    - type: ”performance"
        - scope: “workload”
        - rate: “maximum”
    - type: ”performance"
        - scope: “server”
        - rate: “maximum”
    - type: ”SMART"
        - scope: “disk”
        - rate: “maximum”
    - type: ”metadata"
        - scope: “device-type”
        - rate: “maximum”
    - type: ”age"
        - scope: “device-type”
        - rate: “maximum”

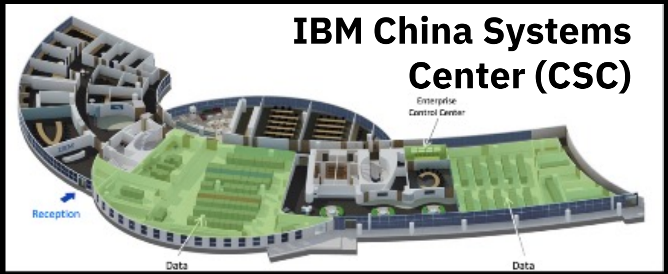
How it works – Adaptive Scheduler (2)

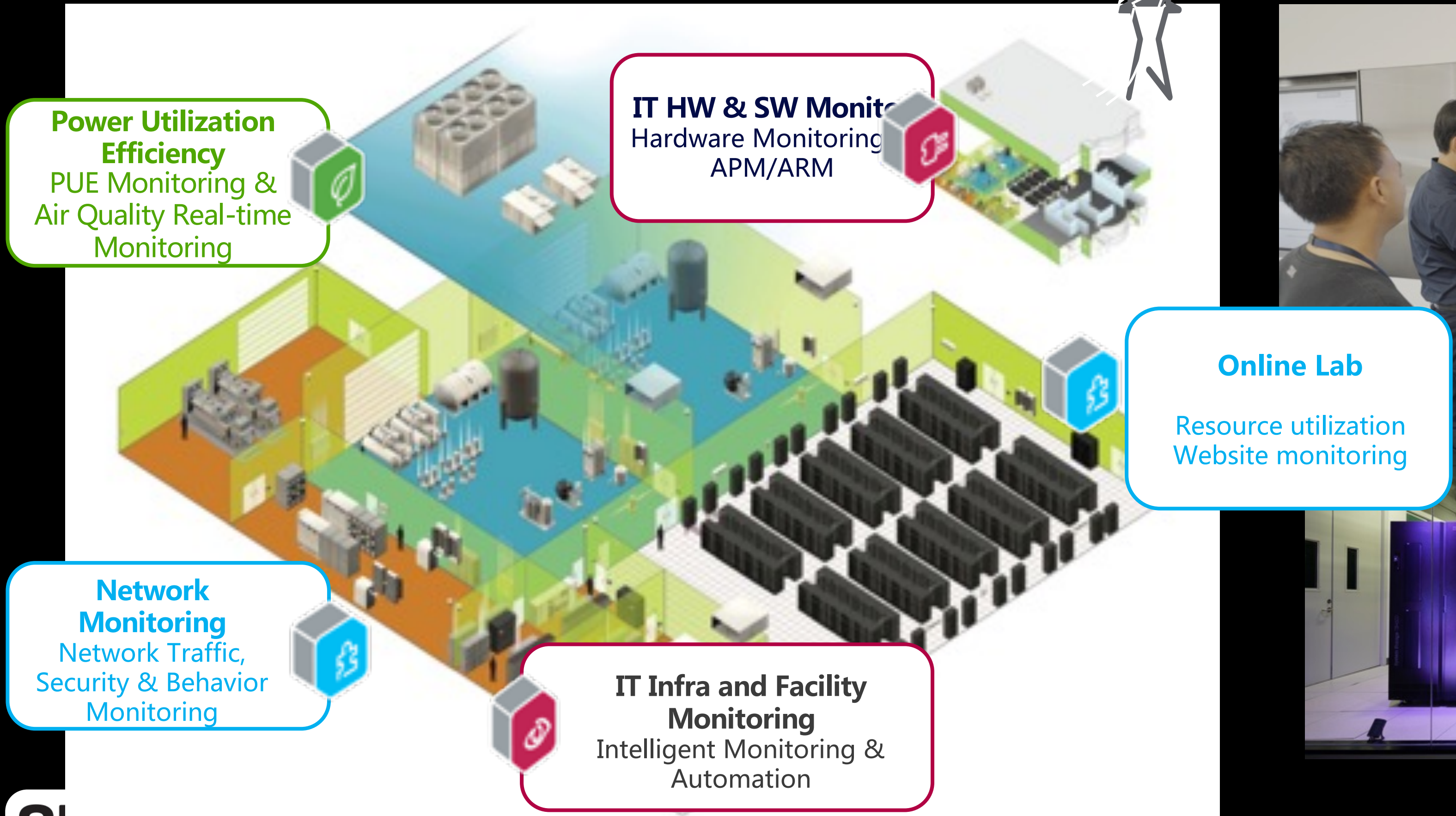
- Detailed approaches:
  - Evaluator:
    - Predict failures with delta data of external benchmark DB, historical data, failure DB
    - Evaluate the model drifting
  - Scheduling Policy Determination
    - Select policies for active sampling and/or prediction model training
  - Schedule jobs based on select policies



Deployed in IBM China Systems Center for daily operations and support visit, demo, testing, PoC, pilot and co-creation

DatacenterDynamics AWARDS





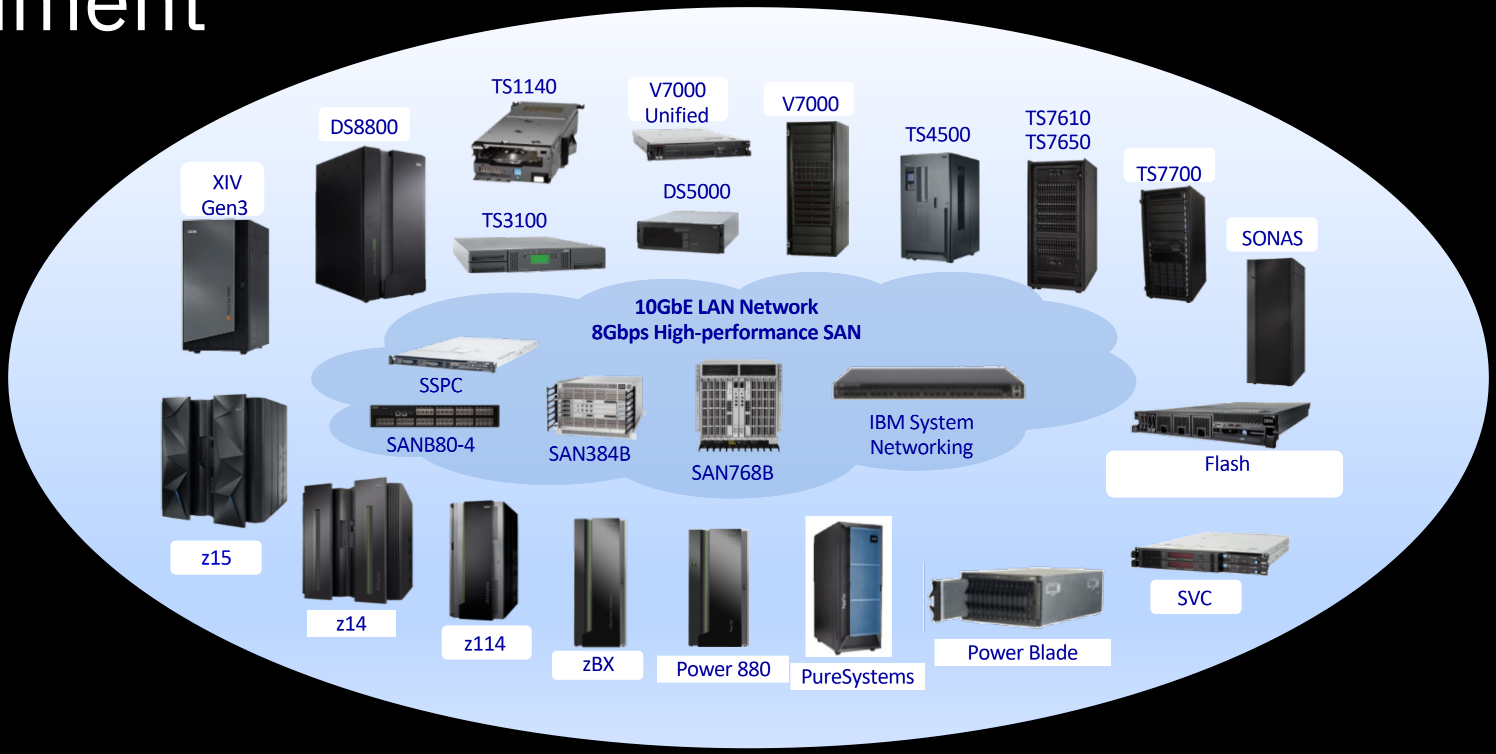
PoC Deployment Environment

Servers

- z15, z14, LinuxOne
- Power9, Power8, Linux on Power
  Power780/770/750
- x86

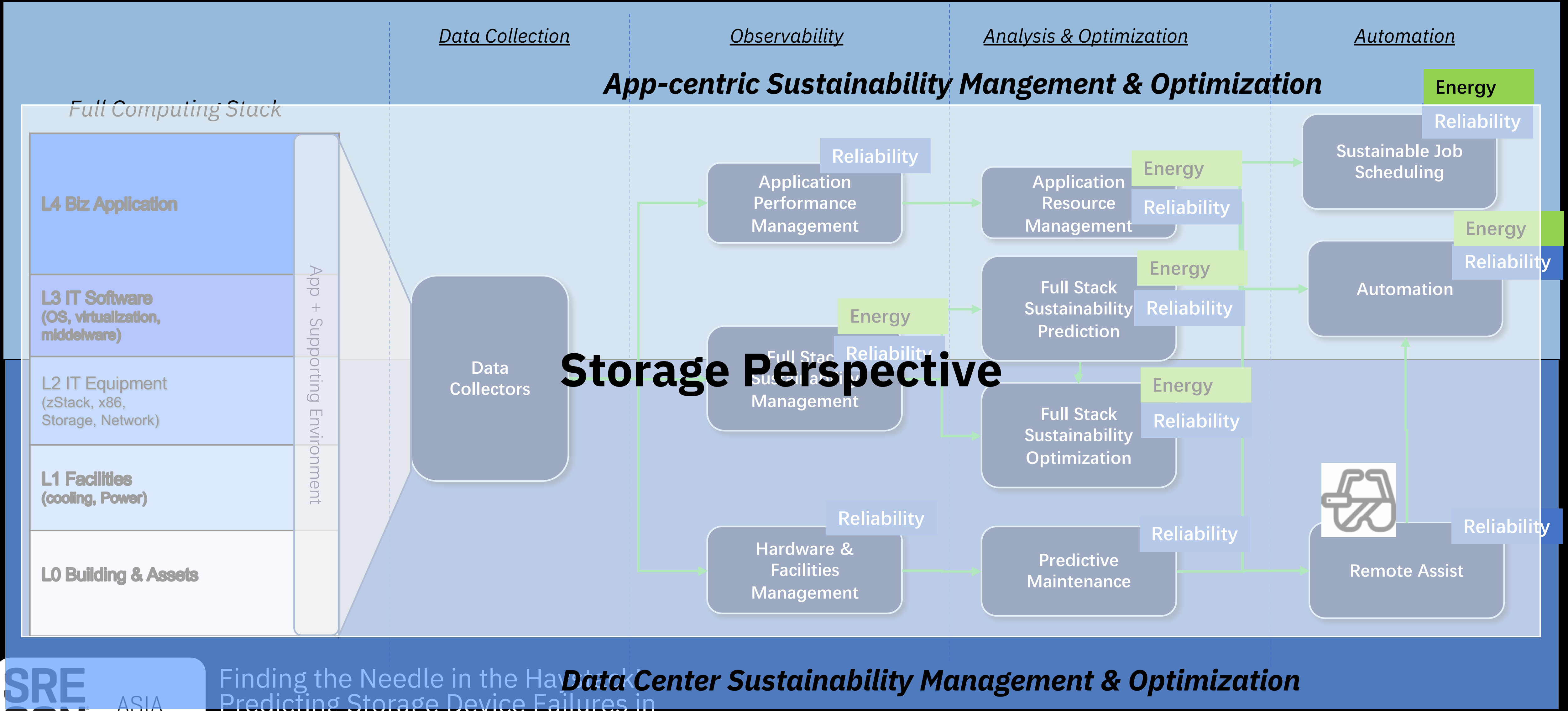
Storages

- DS8886
- Flash9200,
- TS7700, TS4500
- SVC, V7000

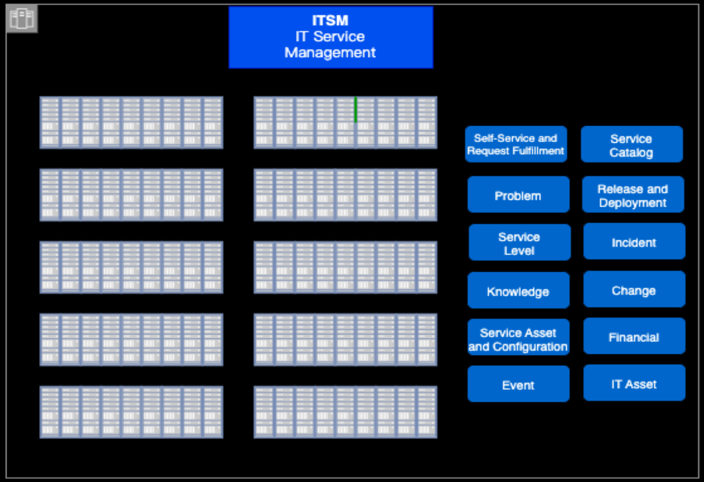


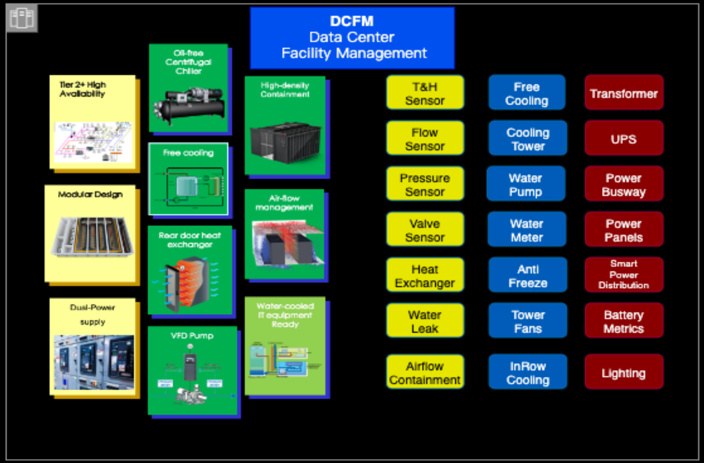
Full-Stack Sustainability Optimization Platform

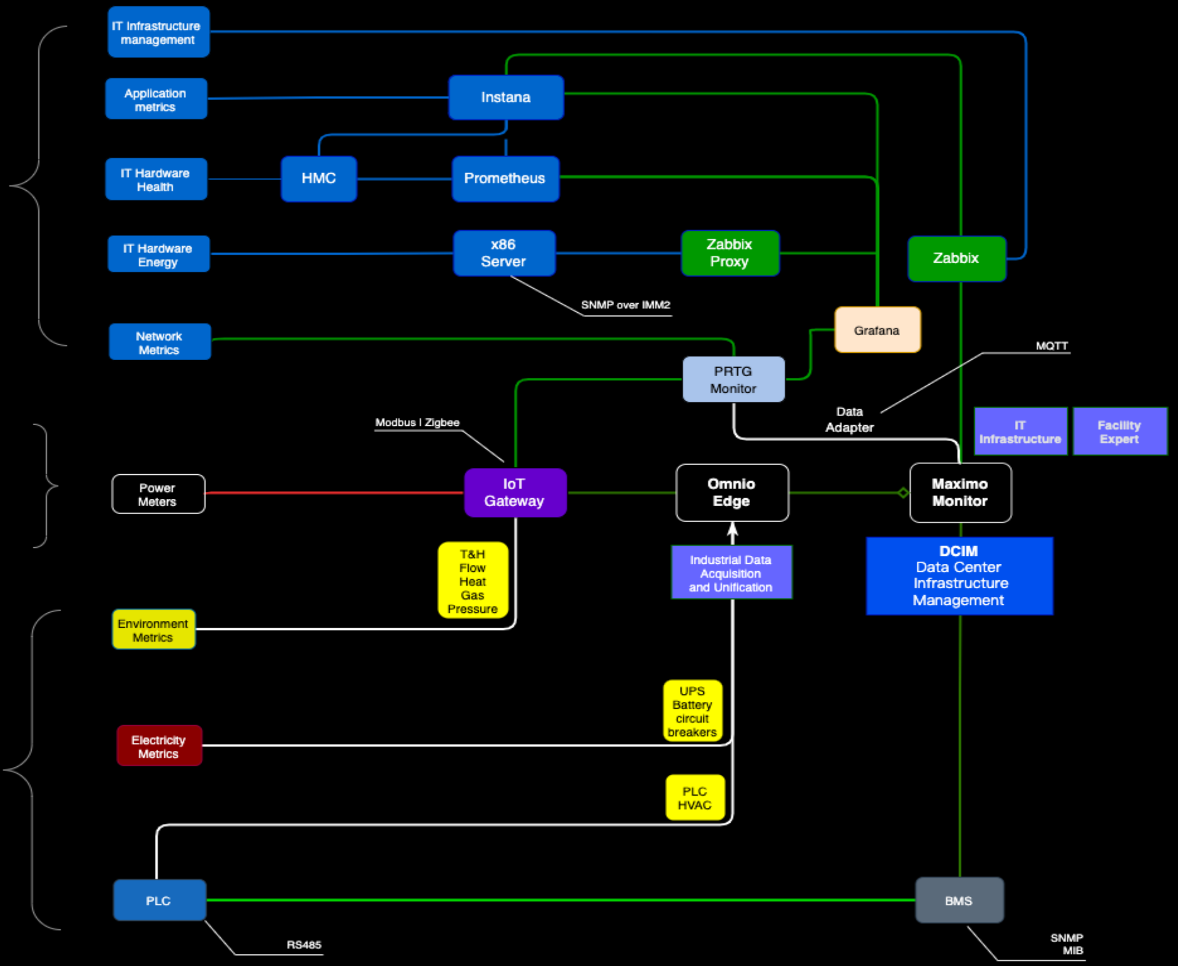
Storage Perspective



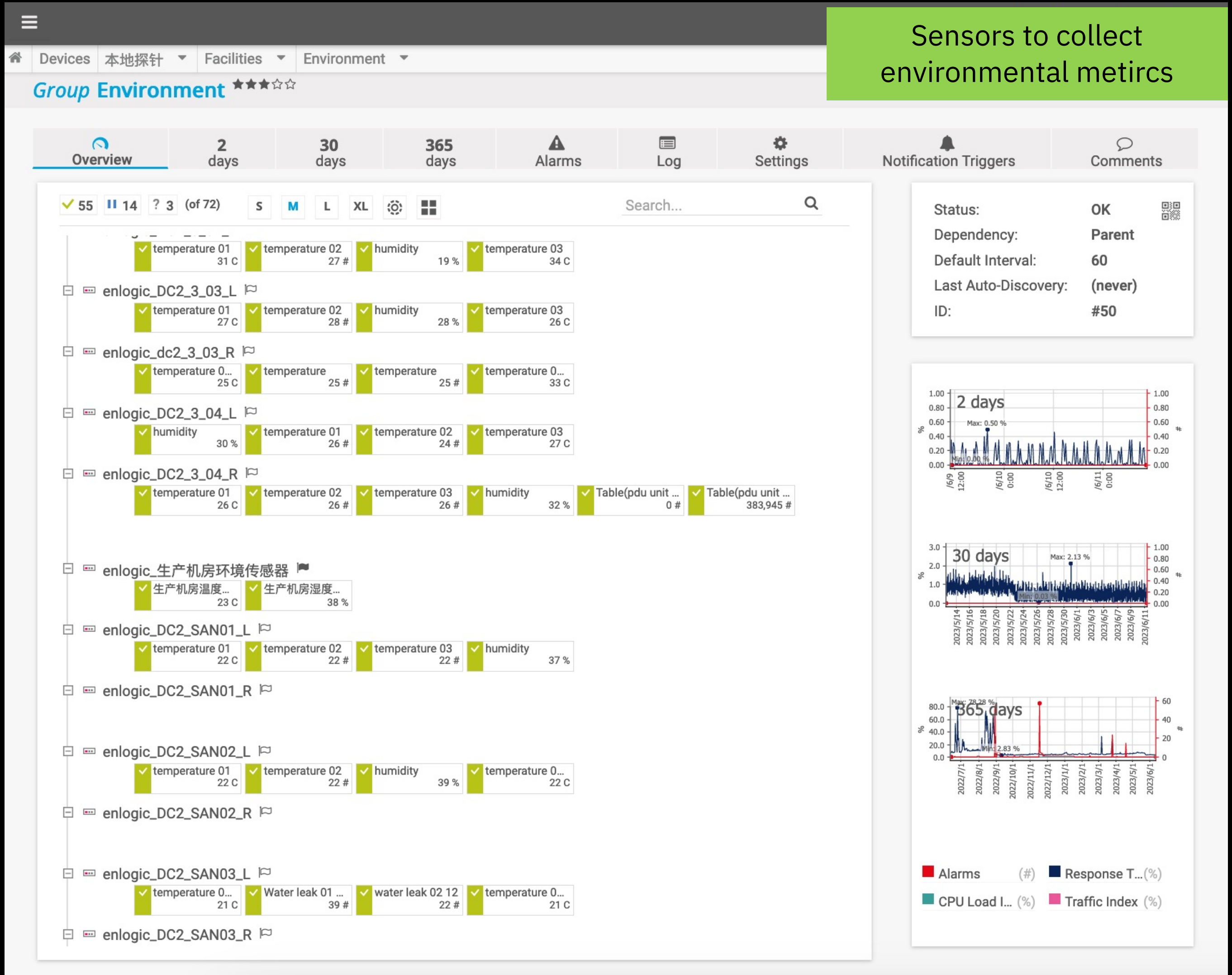
Real-time Monitoring Data Pipeline and Systems for the Whole Stack

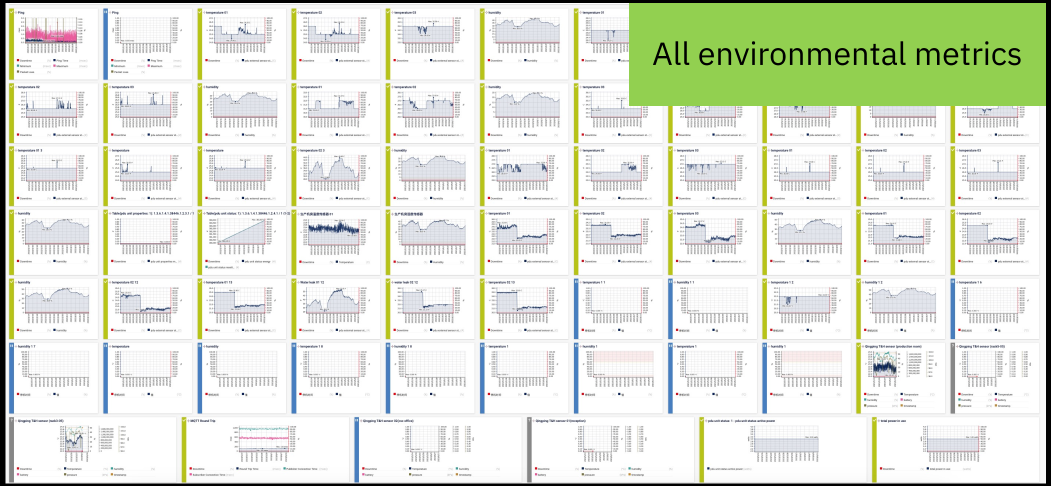




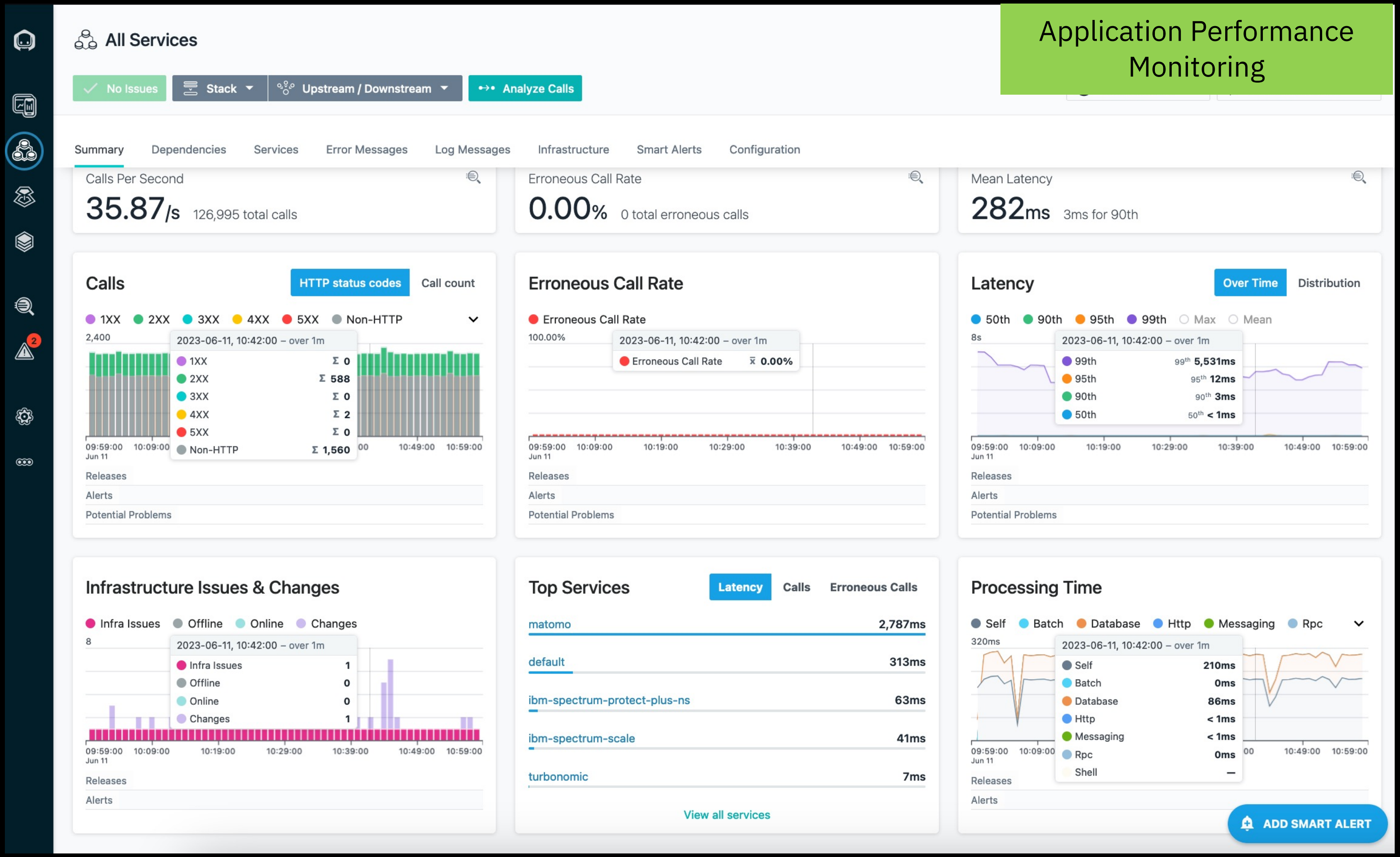


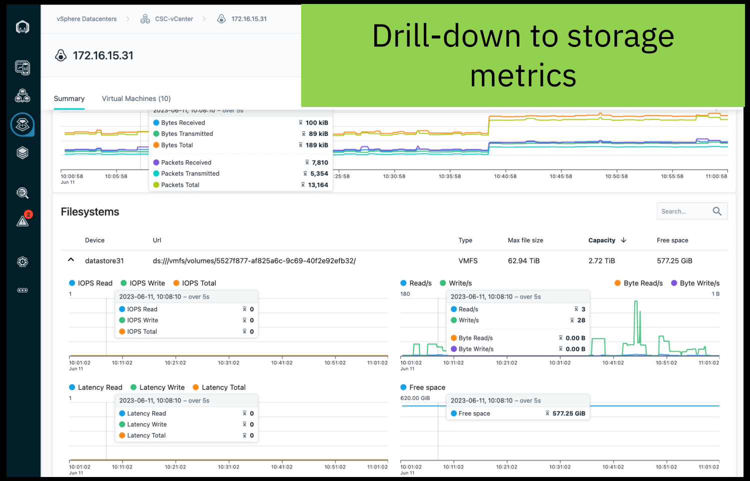
Environmental Metrics Monitoring

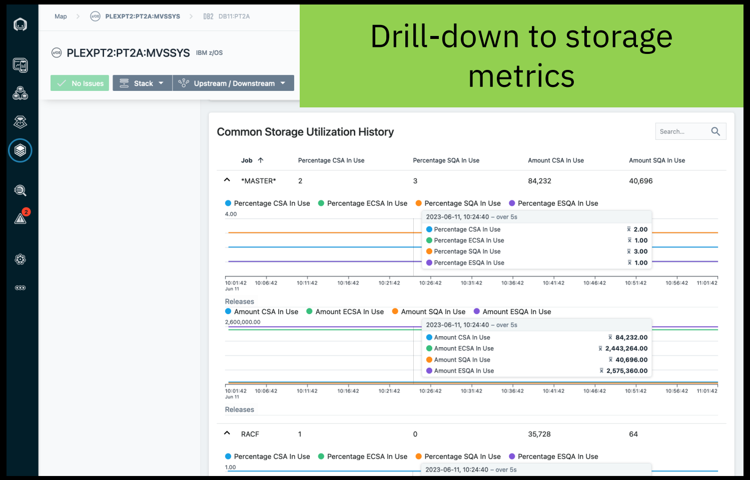




Application Performance Monitoring

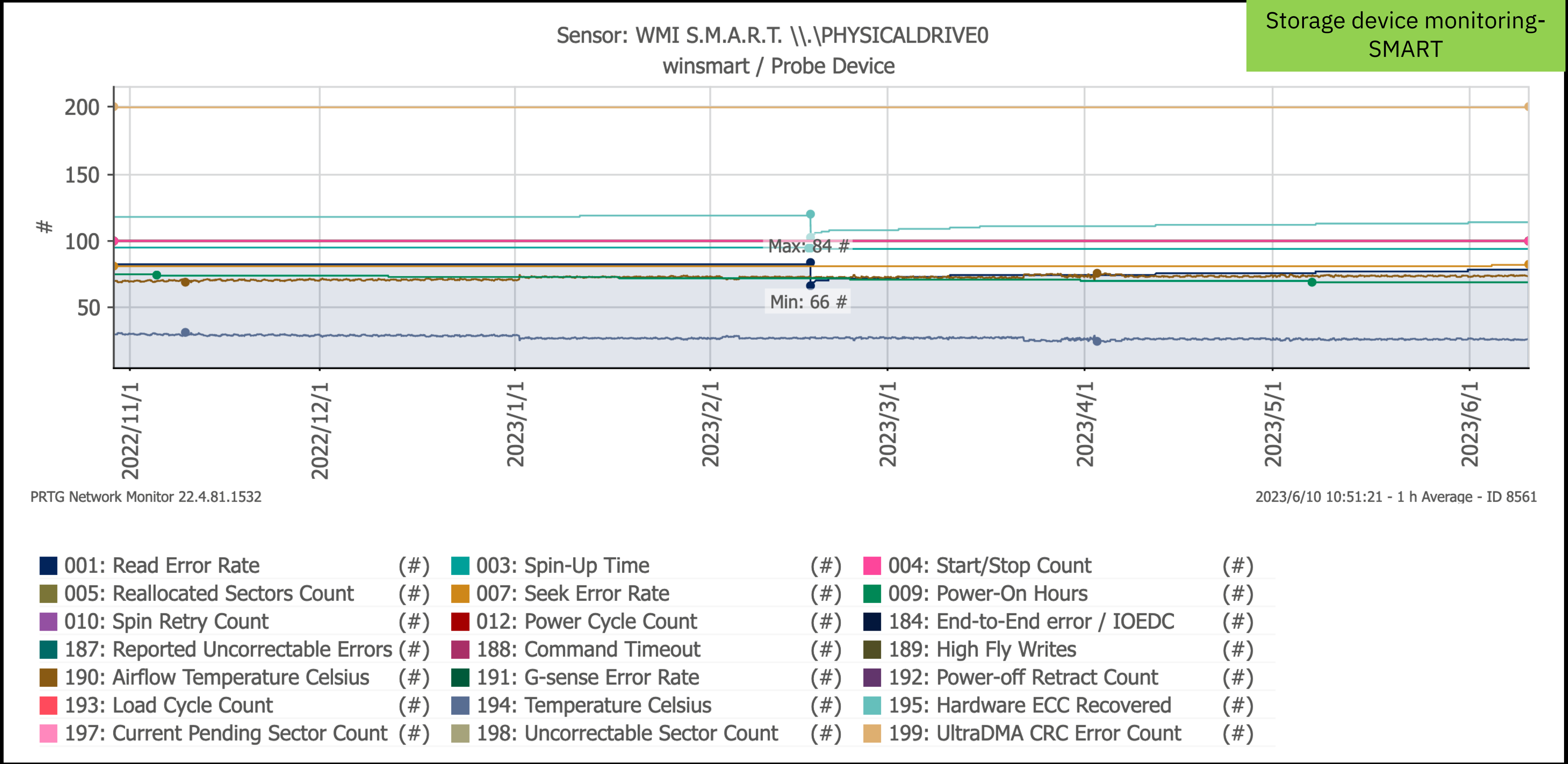




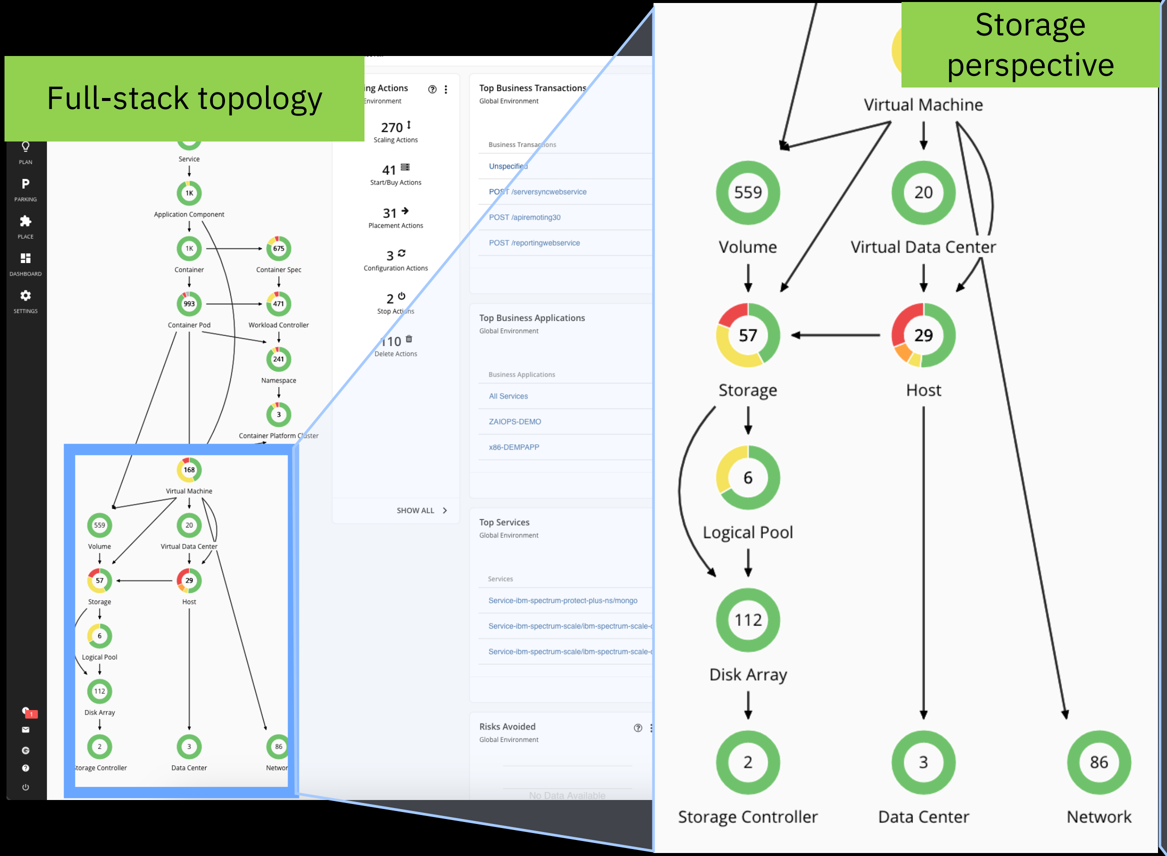


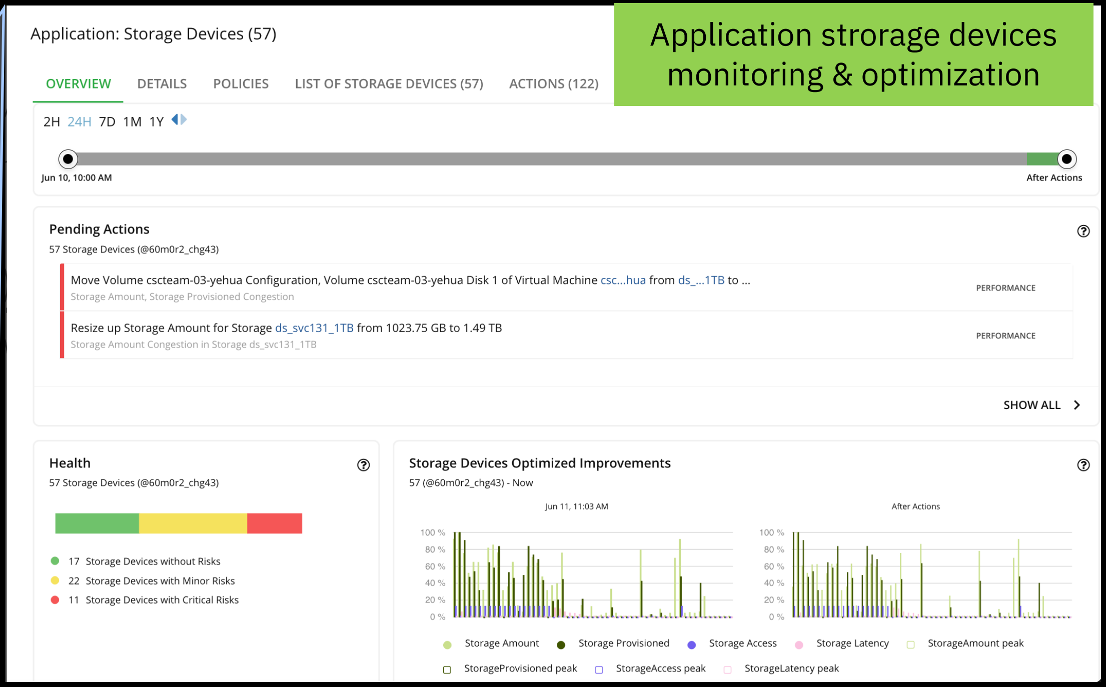
Storage Device Monitoring - SMART

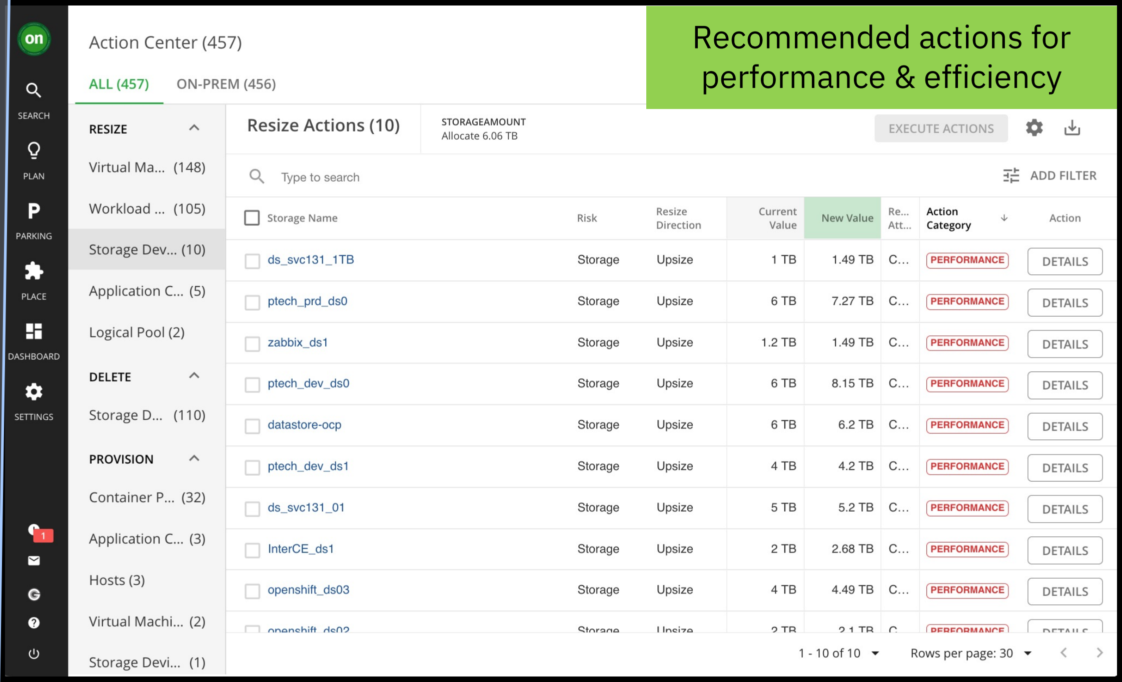
Storage device monitoring- SMART



Storage Resource Management & Optimization







Storage Asset Management with AI & AR

1.EAM

2.IoT

3.AI

Health

Predict

Assist

Asset Lifecycle Mgmt

Real-time Monitor

Analyze Health Status

Predict Failure Risk

Diagnose Failure with AI+AR Assist

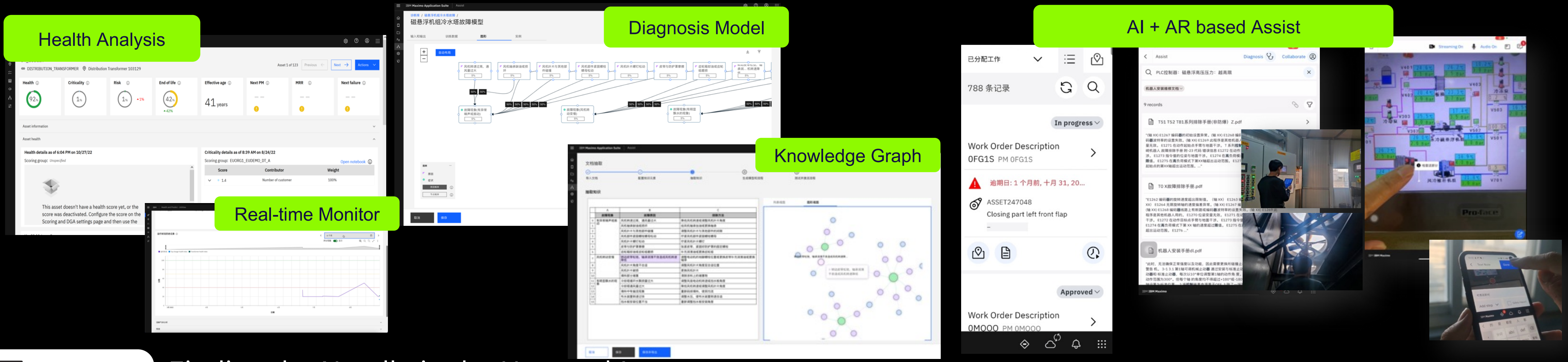
Manage
Keep critical assets operating at maximum efficiency

Asset Monitor
Connectivity enables data aggregation and condition monitoring

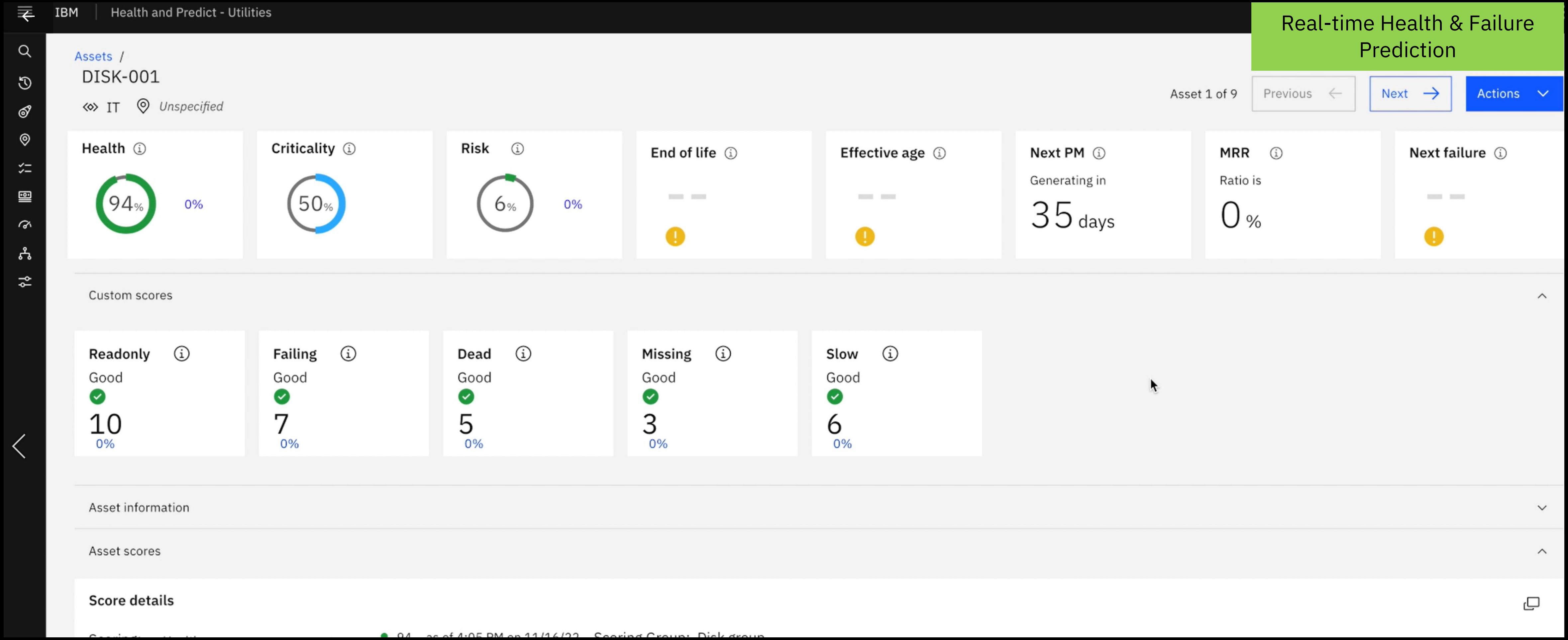
Maximo Health
Analyzes maintenance and IoT data to help prioritize maintenance & replacement activities

Maximo Predict
Analyzes device properties to predict failures and performance degradation

Maximo Assist
Helps identify the right steps to fix an issue given current symptoms and history

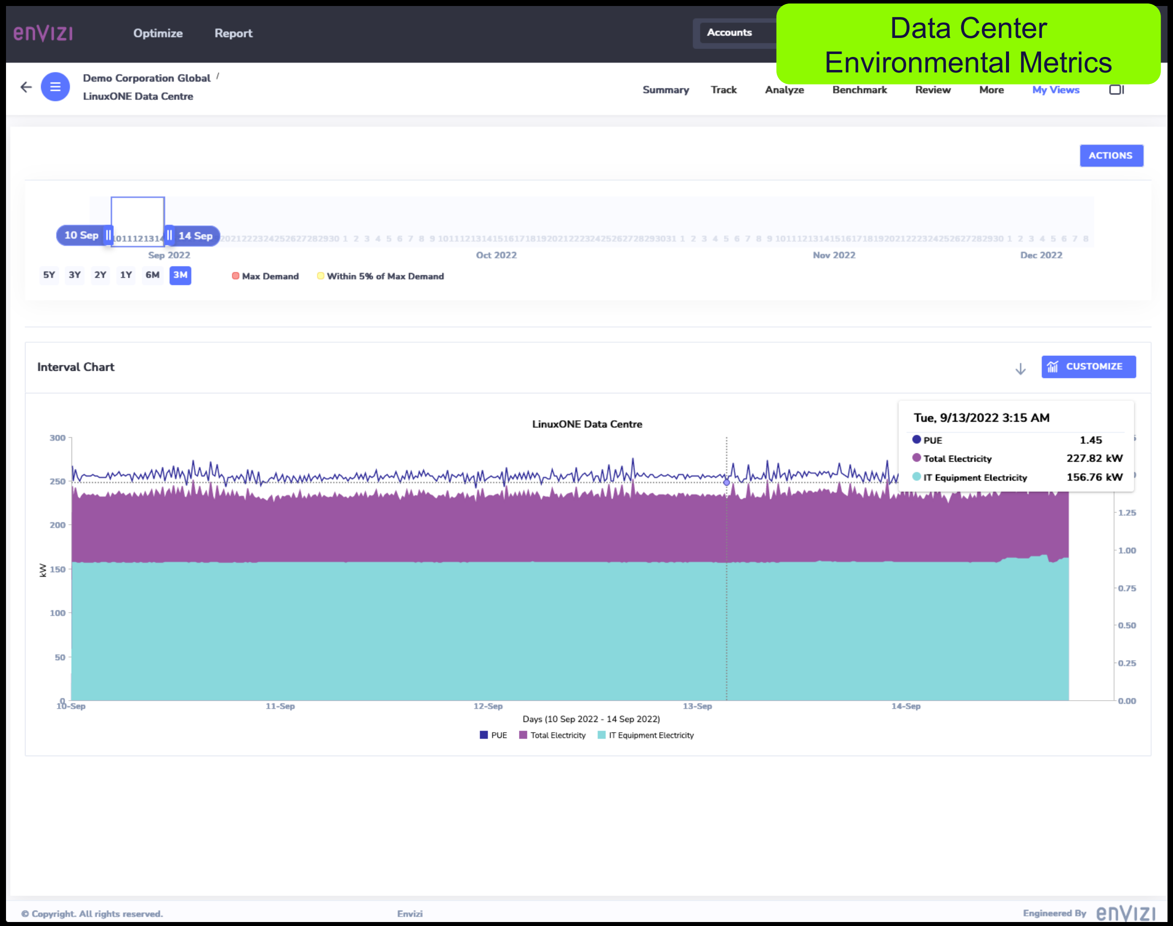


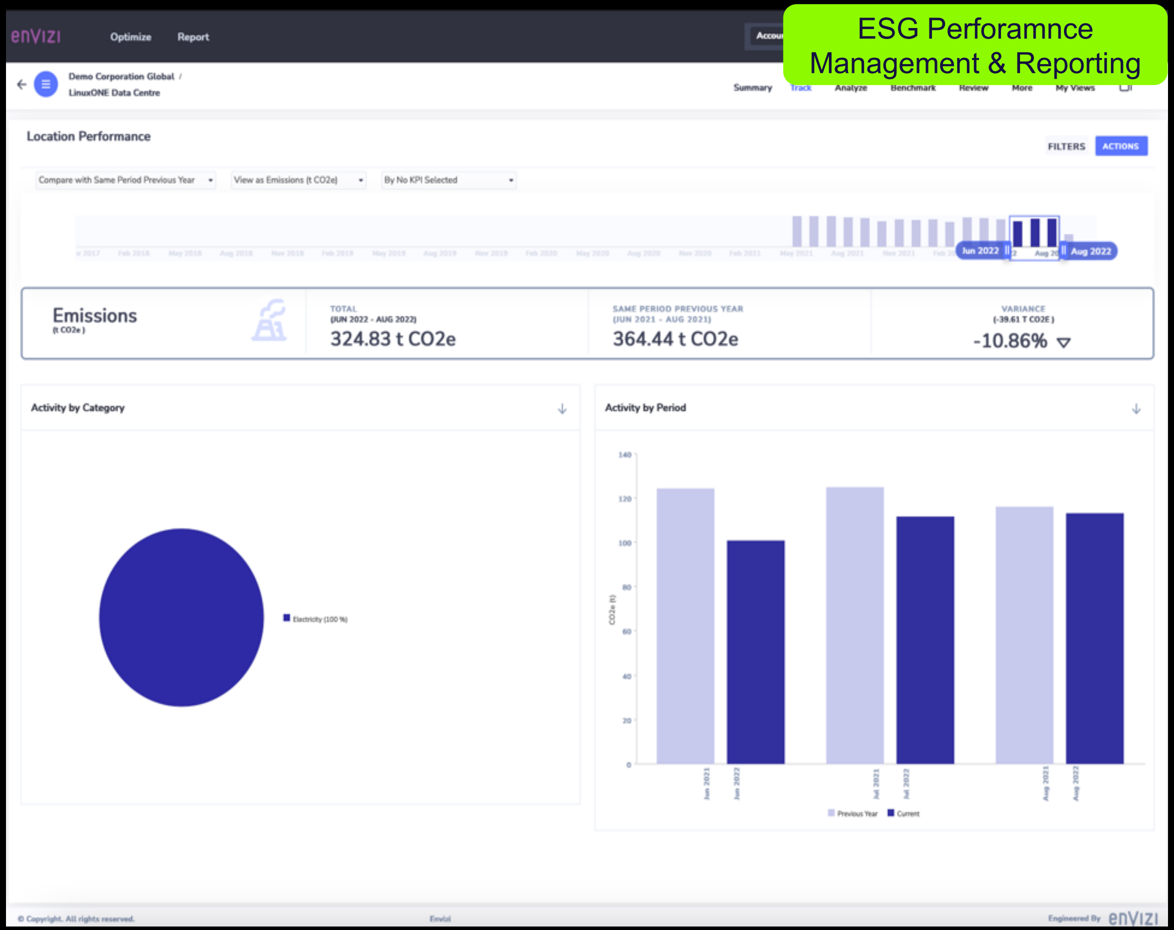
Storage Device Health & Failure Prediction



Storage Power Consumption Monitoring

ESG Performance Management and Reporting





Take-aways

- Health of storage devices are vital for any enterprises and service providers in order to ensure the business continuity and data protection.
- It needs a full-stack approach to manage storage devices from environment, hardware, software, and up to application level.
- Storage device failure prediction should consider and balance the performance, accuracy, sustainability and cost.
- Analysis should be in real-time and continuous optimization.

Acknowledge

We appreciate all the contributors to this Full-stack Sustainability Solution Co-creation and Practices!

Thank You ！

Any questions or comments, please contact Meg (mengfj@cn.ibm.com) !

IBM

Links

Github

For links to public resources, refer to the contact information on the introduction slide.