Acknowledgement(s)

Ivan Schreter, Andrei Mihnea, Oliver Rebholz, Daniel Boos, Werner Thesing, Zora Caklovic, Thomas Peh, Mehul Wagle, Robert Schulze, Christian Lemke, Carsten Thiel, Ami Castiel, Akanksha Meghlan, Rolando Blanco, Muhammad Sharique, Sebastien Seifert, Ami Castiel, Surendra Vishnoi, Jody Glider
Agenda

• SAP DATA Strategy
• Current State of SAP Storage (SAP Cloud)
• Resiliency, Availability and Performance
• Memory Centric Storage (NVM)
• Future Considerations
• Summary
Bi-Model IT

Business Intelligence

Mode 1

Mode 2

SAP Data

SAP Data
SAP DATA

Mode 1
- Block
- File
- Object
- Tables
- Stream
- Queue
- KV

Mode 2

HANA

VORA

SAP Data

Open Source Storages (e.g. PostgreSQL, Mongo, Redis, ...)

Infras as-a-Service (or others e.g. Amazon | Google | Microsoft)
SAP DATA - TODAY
SAP STORAGE TYPES

Big Data

- Hadoop
- Vora

Sap Data

- IQ
- HANA
- ASE
- Postgres

Volume

System of Record

Storage

Files, Objects,

VMs, application
data...

Variety

Value

© 2017 SAP SE or an SAP affiliate company. All rights reserved.
SAP Storage – Historical Data

Capacity growth is 7.7X over 6 year period.

Price Trends 2005-2016

SAP Storage Capacity Trend (Trailing 12 months)
Historical Data & Trend

- Installed base exceeds 150PB: 65% File (NAS), <10% Block (SAN), < 1 PB Object and <20% PB Backup
- Constant growth on File (NAS) and no growth on Block (SAN) - Growth rate for File 20+%, Block 5%, Object (Just started) and Backup 30+% 
- 1200+ unique systems/platforms host bulk of SAP Data.
- Today Storage is consumed (to-date) per application/deployment.
- Installed based on >90K Drives >10K SSDs (most new deployments are SSD)
- Price erosion of 1500% over 11 years with 136% PA.
- Failure rate for HDD – approx 1.6% vs SSD 0.73%.
- Storage Modernization (Today + Tomorrow) is focused on
  - Higher order of resiliency and geo distributed and availability
  - memory centric storage and
  - support new and emerging data types.
  - Storage delivered as a service
SAP DATA: High Level Architecture

Tier-1
- Perf Tier
- Ethernet 100GE

Tier-2
- Capacity Tier
- Ethernet 10GE

Tier-3
- Cloud Integrated

SAP IaaS
- Manila API
- File Store

VM Payloads
- Cinder/Nova
- Block Store

Object Service
- Swift/S3
- Object Store

SaaS
- Cloud App
- Application (e.g. HCP@CF)

PaaS
- Application Platform
- Application (e.g. HCM, B1Cloud, MobileSecure)

Storage Fabric
- T1 (hot)
  - File, Memory
  - >2-way Replication
  - SSD / NVMe

- T2 (warm)
  - File, HDFS
  - 3-way Replication
  - HDD

- T3 (cold)
  - S3, Swift
  - Erasure Coding
  - HDD / Tape / BluRay

© 2017 SAP SE or an SAP affiliate company. All rights reserved.
SAP DATA – Software Defined
Server aligned and Multi-vendor

Distributed Data Store

- Manila API File Store
- Cinder/Nova Block Store
- Swift/S3 Object Store

Storage

Server with NVMe drives

4x100 Gbit

4U 60 drive HDD
RESILIENCY, AVAILABILITY & PERFORMANCE
SAP Customer Facing Data Center footprint

Data Matters
Resilient and Available Everywhere!

- Core Data Center location live or in planning
- Data Center location under evaluation
- Metro setup - allows high-availability setups
- Data Center with long distance disaster recovery capability
Resiliency - Design Target

• Target drive, node, rack failures (and DC – subject to connectivity).

• 3-way vs. 2-way replication MTTDL Model Comparison
  • 10 PB population, 4 TB drives (assumption: first year aggregate population is less than 10 PB)
  • Server MTTF ~100k hours, SSD MTTF ~1M hours

<table>
<thead>
<tr>
<th></th>
<th>MTTDL (Yrs)</th>
<th>Capacity w/o resilency spec</th>
<th>Rack Failure Resilience</th>
<th>Recovery Time</th>
<th>Capacity with Resiliency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dual Parity RAID</td>
<td>&gt;2000</td>
<td>1.35x</td>
<td>No</td>
<td>slow</td>
<td>2.6x</td>
</tr>
<tr>
<td>2-way</td>
<td>9</td>
<td>2x</td>
<td>Yes</td>
<td>fast</td>
<td>2x</td>
</tr>
<tr>
<td>3-way</td>
<td>2695</td>
<td>3x</td>
<td>Yes</td>
<td>Faster</td>
<td>3x</td>
</tr>
<tr>
<td>2way with RAID-5</td>
<td>&gt;10,000</td>
<td>2.2x</td>
<td>Yes</td>
<td>slow</td>
<td>2.2x</td>
</tr>
</tbody>
</table>

• Min. 2-way replication with RAID-5/6 or underneath initially. With flash prices going down, 3-way will be affordable in 2018 and beyond. Considerations – replication based on data type, Flash+HDD, # of availability zones, 2017 vs 2020
• Replication is a necessary evil (well understood today). That drives cost/bit economics and thus sourcing and storage architecture.
• Near Term - Sync Replication within DC with Async/Mirror for across DC
Performance: Addressing Latency & throughput

- Limitations of OS Network Stack:
  - **Cannot saturate NIC** for small requests & establish multiple connections
    - Context switches, data copies, OS-locks
    - Packet processing on "wrong" core
    - OS buffer per connection
    - Practical limit: ~20K TCP/IP connections/system

- Mitigation: **User-Space Networking**
  - Direct access to NIC
    - Supported even in cloud
  - Pre-allocated network buffers, zero-copy usage
  - Per-thread queues, no context switches
  - Minimal connection overhead (~300B)
    - 10M+ active TCP/IP connections possible

- Alternative: RDMA & Co.
  - Special HW – limited applicability
  - External communication still TCP/IP
  - ...but fits into our model
User-Space Local I/O

NVMe 4k IOPS on Single Core*

- SPDK Library
  - NVMe driver in user space
  - Analogous to Network I/O
    - Per-thread queues
  - Millions of IOPS on single core

*) Source [here](#), using Intel P6000 SSDs. Intel Optane performance with up to ~5.6 MIOPS even better.
Revisiting CAP Theorem¹

• Distributed systems should be:
  • Consistent (return most recent data)
  • Available (every request receives a response)
  • Partition-tolerant (system works despite network failures)

The catch: impossible to fulfill all three

Traditional scale-up databases
  • C only (ACID transactional)

Distributed databases and “NoSQL” data stores²
  • CP: prefer consistency over availability at partition
  • AP: prefer availability during partitions over consistency

---

¹ See also PACELC theorem, which addresses some of the over-simplifications in CAP theorem
² The classification is very rough, see for example stop calling databases CP or AP
Elastic Scalability, Availability and Resilience

- **Existing nodes**
  - Many virtual partition replicas evenly assigned
  - Every node has \( \frac{1}{n} \)th of the data

- **New node**
  - Assigned fair share of virtual partition replicas
  - Whole partition replicas moved – simple
  - Rebalancing uses network I/O optimally

- **Remove node**
  - Analogous – move partition replicas to surviving nodes
Upholding SLAs*: High Availability – Large Outage

<table>
<thead>
<tr>
<th>Node 1</th>
<th>Node 2</th>
<th>Node 3</th>
<th>Node 4</th>
<th>Node 5</th>
<th>Node 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>P1</td>
<td>P1</td>
<td>P2</td>
<td>P2</td>
<td>P2</td>
</tr>
<tr>
<td>P3</td>
<td>P3</td>
<td>P3</td>
<td>P4</td>
<td>P4</td>
<td>P3</td>
</tr>
<tr>
<td>P5</td>
<td>P5</td>
<td>P5</td>
<td>P6</td>
<td>P5</td>
<td>P5</td>
</tr>
<tr>
<td>P7</td>
<td>P7</td>
<td>P7</td>
<td>P7</td>
<td>P8</td>
<td>P8</td>
</tr>
<tr>
<td>P9</td>
<td>P9</td>
<td>P9</td>
<td>P11</td>
<td>P10</td>
<td>P9</td>
</tr>
<tr>
<td>P11</td>
<td>P11</td>
<td>P11</td>
<td>P12</td>
<td>P11</td>
<td>P11</td>
</tr>
<tr>
<td>P12</td>
<td>P12</td>
<td>P12</td>
<td>P12</td>
<td>P12</td>
<td>P12</td>
</tr>
</tbody>
</table>

* Upholding SLAs: High Availability – Large Outage

Node 1
- P1
- P3
- P5
- P7
- P9
- P11

Node 2
- P1
- P3
- P6
- P7
- P9
- P12

Node 3
- P1
- P4
- P6
- P7
- P10
- P12

Node 4
- P2
- P4
- P6
- P8
- P10
- P12

Node 5
- P2
- P4
- P5
- P8
- P10
- P11

Node 6
- P2
- P3
- P5
- P8
- P9
- P11
Platform Evolution 2015-2017
Addressing Availability

- Media - Flash, SCM and NVMe. 8-32 TB
- Networking – 100G Ethernet
- Performance:
  - Throughput - 20-30 GB/s (limited by network)
  - Latency: Needs rethinking
- Cost: Approach cost of NVMe drives
- Resiliency: Failure domain – 1 controller + 4-6 NVMe drives.
- Everything is Software Defined
- Enable new data types
MEMORY CENTRIC STORAGE
Memory & Storage Hierarchy

<table>
<thead>
<tr>
<th>Hierarchy</th>
<th>ns</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1 Cache</td>
<td>0.5</td>
<td>1 sec</td>
</tr>
<tr>
<td>L2 Cache</td>
<td>7</td>
<td>14 sec</td>
</tr>
<tr>
<td>L3 Cache</td>
<td>20</td>
<td>40 sec</td>
</tr>
<tr>
<td>Main Mem</td>
<td>100</td>
<td>3.3 min</td>
</tr>
<tr>
<td>NVRAM(*)</td>
<td>1000</td>
<td>33.3 min</td>
</tr>
<tr>
<td>NVRAM (IO)</td>
<td>10000</td>
<td>5.6 hours</td>
</tr>
<tr>
<td>Flash</td>
<td>50000</td>
<td>1.2 days</td>
</tr>
<tr>
<td>LAN</td>
<td>500000</td>
<td>11.6 days</td>
</tr>
<tr>
<td>WAN</td>
<td>2000000</td>
<td>1.5 month</td>
</tr>
<tr>
<td>Disk</td>
<td>20000000</td>
<td>15.4 month</td>
</tr>
<tr>
<td>Round Trip</td>
<td>150000000</td>
<td>9.6 years</td>
</tr>
</tbody>
</table>

* - Projected

Faster Storage vs Slower memory
PERSISTENT MEMORY & HANA

Opportunities

• Increased scalability
  o Larger memory modules means more memory available per server

• Significant cost savings
  o PM is cheaper than DRAM

• Improved recovery

Challenges

• Higher (than DRAM) latency impacting performance

• New technology, standards still evolving…
  o Means slow, phased implementation with increased complexity and uncertain timelines

HANA and in general in-memory databases/application consume larger memory
Business application workload
MEMORY CENTRIC STORAGE: HANA AND 3D XPOINT

SAP HANA: Memory Architecture

Main Store is the perfect fit for Intel DIMM based on 3D XPoint™ technology!

Main store contains ~95% of the data in highly compressed format.

<table>
<thead>
<tr>
<th>Technology differentiators</th>
<th>Why Main is well suited</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large capacity</td>
<td>Since 95% of data contained in main, HANA can scale up to larger datasets due to increased memory capacities</td>
</tr>
<tr>
<td>Persistence</td>
<td>Avoid loading data from storage and reduce downtime</td>
</tr>
<tr>
<td>Higher latencies</td>
<td>References are in form of scans. Hardware and software prefetchers can hide latencies for such reference patterns</td>
</tr>
</tbody>
</table>
SAP HANA adoption of NVM – Architecture

Large structures, read mostly
NVM integration in the relational engine

Main Column Fragment

Main attribute

NVM Filesystem
Identify the block within the NVM filesystem

Attribute load
- Construct DRAM anchors
- Use offsets
- Point to NVM
- No deserialization of large structures

Offsets
NVM Block
(padded substructures)

Index vector
Dictionary
Inverted index
SAP HANA focus: SNIA programming model for use with memory-like NVM

Intel DIMM based on 3D XPoint™ technology is fully aligned with the SNIA programming models for NVM and provides support for Block, File, Volume, and Persistent Memory (PM) File mode.

**Persistent Memory Mode (SNIA)**

**Use with memory-like NVM**

**NVM.PM.VOLUME Mode**
- Software abstraction to OS components for Persistent Memory (PM) hardware
- List of physical address ranges for each PM volume
- Thin provisioning management

**NVM.PM.FILE Mode**
- Describes the behavior for applications accessing persistent memory
- Discovery and use of atomic write features
- Mapping PM files (or subsets of files) to virtual memory addresses
- Syncing portions of PM files to the persistence domain

SAP HANA focus is on leveraging memory mapping in NVM.PM.FILE mode which enables direct access to persistent memory using CPU instructions.
3D Xpoint – Applicability for HANA

**Pro**
- Significant improvements in database restart time (>5x)
  - No need anymore to load data, Main Store is now persistent as it is in 3D XPoint™ PM instead of DRAM
- Increased memory capacity at a lower cost
  - Potentially significant cost benefits in several areas

**Con**
- New approach has two redundant persistent copies
  - Dual-writes to 3D XPoint™ PM and “SAN”/Disks
    - Necessary for Backup and Recovery (Protect against DIMM failures)
    - Media and access methods are still evolving
  - When to operate directly on “PM” vs “DRAM” ? (Data storage vs intermediate data handling)
    - Both storages co-exist in DDR4 form factor, but the capacity ratio is unknown
Simulation Results (PMEM latencies)
FUTURE CONSIDERATIONS
Platform Roadmap

CPU

CPU

CPU

CPU

2017

4socket (104 cores)
Memory: 6TB DRAM
Storage: 144 TB

1.5X
1.5-3x
8+x

2020

4socket (144 cores)
Memory: 3 TB DRAM + >6 TB (NVM)
Storage: (1PB) 3D xPOINT + 3D NAND

© 2017 SAP SE or an SAP affiliate company. All rights reserved.

PUBLIC
3D Xpoint – Vertical NUMA

TODAY

67ns

Socket

Socket

Socket

Socket

Flash

120ns

Socket

Socket

Socket

Socket

TOMORROW

>250ns

Socket

Socket

Socket

Socket

SCM

Flash

DRAM

DRAM

DRAM

DRAM

© 2017 SAP SE or an SAP affiliate company. All rights reserved. | PUBLIC
Other Considerations

• Data Protection (Regional, Regulatory)
• Leverage Public Cloud Storage (AWS, GCP, Azure...)


Infrastructure Technology Roadmap
Compute, Network and Storage – All Software Defined

**Compute: Software Defined**
- 1986-2016: CPU Era - MHz & Cores
- 2017 – Future: Dawn of GPU, TPU

**Network: Software Defined**
- Ethernet (10G to 100G to 400G)
- NVM drives Rack Network (NV/MoF)
- Resilient Storage Drives DC Networking
- Multi-cloud drives SD-WAN

**Memory/Storage: Software Defined**
- 2018: NVRAM Tier
- 2022: Flash cost 2x of HDD

**Disruptors:**
- Storage and Micro-services drive Network
- Non-volatile memory and data always available everywhere

**Software Defined Datacenter**

<table>
<thead>
<tr>
<th>2017 Feature</th>
<th>2020 Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>768 Cores</td>
<td>1280 Cores</td>
</tr>
<tr>
<td>30720 GPU (CUDA) Cores</td>
<td>1E+05 GPU (CUDA) Cores</td>
</tr>
<tr>
<td>12 FPGA LUT (M)</td>
<td>768 FPGA LUT (M)</td>
</tr>
<tr>
<td>128 DRAM (TB)</td>
<td>48 DRAM (TB)</td>
</tr>
<tr>
<td>192 NVRAM (TB)</td>
<td>192 NVRAM (TB)</td>
</tr>
<tr>
<td>240 Flash (TB)</td>
<td>2304 Flash (TB)</td>
</tr>
<tr>
<td>100 Network/Node (Gb/s)</td>
<td>200 Network/Node (Gb/s)</td>
</tr>
<tr>
<td>3200 Bisection BW (Gb/s)</td>
<td>12800 Bisection BW (Gb/s)</td>
</tr>
<tr>
<td>800 Uplink BW (Gb/s)</td>
<td>3200 Uplink BW (Gb/s)</td>
</tr>
</tbody>
</table>

**Challenge:** Finding the right balance of price, perf and resiliency for cloud
SAP DATA

Mode 1
- Block
- File
- Object
- Tables

Mode 2
- Stream
- Queue
- KV

HANA

VORA

DATA = STORAGE + (PM)-MEMORY
SAP DATA

Renu Raman
SAP