Windows File and Storage Directions

Surendra Verma
Partner Development Manager
Windows Storage & File Systems
Landscape

- Pooled Resources
- Self-Service
- Elastic
- Virtualized & Usage-Based
- Highly Available
Industry trends and challenges

- Services influencing hardware design
- Industry standard “commodity” parts to build very high-scale, performant systems
- New apps/services
- Device proliferation
- Data explosion
- Cloud computing
- New technologies exploit these trends to address specific needs
  e.g., “big data” solutions
- Use of greater numbers of less reliable parts and building resiliency atop is an obvious trend
Opportunities and Challenges

Bring benefits of these ideas and trends to traditional OS and applications

Drives deep changes throughout the OS stack. A significant investment

Start by understanding key attributes and how they ought to be applied
### Key Attributes

| Design for “services” and the data center | • Think in terms of racks of servers with enclosures full of storage elements (disks, SSDs etc) with storage served over the network to a set of compute nodes over multiple protocols |
| End-to-end thinking around failures | • Disks, enclosures, power supplies, cables, controllers, CPU/machine, network, racks, datacenters |
| Resiliency in software is key | • Redundancy in both data storage and access paths  
• Verification, correction, proactive scanning  
• Not surprising and not new, but significant |
| Scale and cost | • Data explosion requires massive scale without proportionate cost and management increase |
Key impact to Windows - Flip the assumption on reliability

- Expect data loss and corruption and be defensive
- Traditional approach to rebuilding in RAID doesn’t scale
- Very large scale of deployment requires scalable data management
- Closer integration with hardware matters
Innovations in Windows

**NEW TECHNOLOGIES**

**ReFS**
Resilient, highly available and scalable file system

**Storage Spaces**
Reliable, virtualized pooling of storage, shared across multiple machines

**BENEFITS**

Redundant storage, pooled across a cluster of machines

Highly resilient and scalable file system layered on top to take advantage of the pooled redundant storage to provide end-end data integrity

Common volume namespace across the cluster of machines enables easy deployment

SMB Direct for efficient access of data across nodes
Resiliency in software
Resiliency to media failures

- Redundant Storage via Mirror (2 & 3 way) and Parity Storage Spaces
- Transparent healing of bad blocks and failed media in Spaces
- Rapid recovery via per-pool hot spares that kick in automatically
Resiliency to node and path failures

- Shared Storage Space Pool available across nodes
- Multi-path I/O, Enclosure aware allocation and repair
Resiliency to Media Errors

- **Torn Writes can not happen**
  - ReFS never writes metadata in place, avoiding this problem

- **Media errors detected and healed**
  - ReFS Integrity Streams checksum user data and never write in place
  - Checksum is validated on all reads
  - ReFS queries Storage Space for multiple copies, picks the authoritatively correct copy
  - Storage Spaces heals the bad copies from the known good copy
  - API to inspect and heal copies exposed up the stack and leveraged by other components (for example deduplication running on NTFS)
Resiliency to Latent Media Errors
Detected and proactively corrected – keeps redundancy levels intact

With NTFS, Scrubber verifies all copies can be read.

On media failure, NTFS will get an alternate copy of the data.

With ReFS, Scrubber verifies checksums on all copies; then auto-corrects as needed.

1. Checksum is inspected on all data copies
2. If corruption is detected, trigger auto-correct and return a healthy alternate copy of the data
Resiliency to loss of all copies

- ReFS continues to make healthy parts available
- Allows recovery of the healthy subset of namespace online without making all data unavailable
Resiliency to All Failures except Media

• All persistent formats are self-described
• A set of media can be instantiated on machine(s) different from the original
Design for Services and Datacenter
REPRESENTATIVE DEPLOYABLE ELEMENT (4-NODE)

RAID Inc. 60-Bay EBODs
3TB SAS NL HDDs

RAID Inc. 60-Bay EBODs
600GB SAS 15K HDDs

RAID Inc. 60-Bay EBODs
400GB SAS SSDs

720TB High-Cap
144TB Mid-Perf
96TB High-Perf
## Recap of Key Attributes

<table>
<thead>
<tr>
<th>Design for services and the data center</th>
<th>Clustered, flexibly provisioned, virtualized deployment</th>
</tr>
</thead>
<tbody>
<tr>
<td>End-to-end thinking around failures</td>
<td>Resilient to failures in disks, enclosures, power supplies, cables, controllers, CPU/machine, network, racks, datacenters</td>
</tr>
<tr>
<td>Resiliency in software is key</td>
<td>Resiliency is implemented in Software at various layers of the stack leveraging investments in Storage Spaces, ReFS, Clustering, SMB Direct and more</td>
</tr>
<tr>
<td>Scale and cost</td>
<td>New Scalable File System and Storage Virtualization, Scalable management via Powershell Scripting and support for Industry Standard Commodity parts</td>
</tr>
</tbody>
</table>