Disclaimer

• This presentation and/or accompanying oral statements by Samsung representatives collectively, the “Presentation”) is intended to provide information concerning the SSD and memory industry and Samsung Electronics Co., Ltd. and certain affiliates (collectively, “Samsung”). While Samsung strives to provide information that is accurate and up-to-date, this Presentation may nonetheless contain inaccuracies or omissions. As a consequence, Samsung does not in any way guarantee the accuracy or completeness of the information provided in this Presentation.

• This Presentation may include forward-looking statements, including, but not limited to, statements about any matter that is not a historical fact; statements regarding Samsung’s intentions, beliefs or current expectations concerning, among other things, market prospects, technological developments, growth, strategies, and the industry in which Samsung operates; and statements regarding products or features that are still in development. By their nature, forward-looking statements involve risks and uncertainties, because they relate to events and depend on circumstances that may or may not occur in the future. Samsung cautions you that forward looking statements are not guarantees of future performance and that the actual developments of Samsung, the market, or industry in which Samsung operates may differ materially from those made or suggested by the forward-looking statements in this Presentation. In addition, even if such forward-looking statements are shown to be accurate, those developments may not be indicative of developments in future periods.
Looking at TCO for ODSA Devices

Accelerators can contribute to the vertical scaling of servers. Samsung has experimented with methods to offload CPUs.

- CPUs normally saturate at ~4-6 SSDs
- A KV SSD was created that would assemble blocks into a value (small amount of compute)
- Scaling worked for up to 18 SSDs
- CPUs freed up to do more work

* Testing was done on a server with 2 x Intel Xeon E5-2600 v5 servers with 384 GB of DRAM, and 18 PM983 (in block or KV mode) SSDs
** Workload: 4KB uniform random writes
Looking at TCO for ODSA Devices

CPU offload and acceleration are already in use.

Accelerator Chips Types

- SmartNICs – Offload network functions, encryption
- GPU – Great for AI/ML, but not efficient
- FPGA – Flexible, but limited clock speed
- ASICs – Expensive to design and manufacture

Chiplets offer an opportunity to save money in the development of accelerator chips.

However adding accelerators will obviously increase cost. So how can we determine the advantages of using them?
ODSA Storage Benefits

There are 2 ways to justify the additional costs associated with compute near storage

<table>
<thead>
<tr>
<th>Better Performance</th>
<th>Better TCO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Here the value is that better performance allows the business to operate differently/at higher scale.</td>
<td>Here the value is that system architectures can be changed achieve a better overall TCO.</td>
</tr>
</tbody>
</table>

Example: A database server can normally handle N transactions/second. With accelerated storage it can now handle 3N, allowing the system to handle more workload.

Example: An analytic deployment uses X servers with Y TB of storage each. Due to CPU offload, each server can handle 3Y TB of storage with no loss of relative performance.

Pros:
- Allows access to performance levels unattainable through normal technology
- Easy to calculate the value of this
- Does not need to have better performance

Cons:
- Hard to calculate the value of additional performance
- Generally costs are higher here
- Must have a good understanding of baseline architecture

Not mutually exclusive!
## TCO Example

An accelerated storage platform can offer substantial TCO savings.

### Goal: Store and process 1 PB of data

<table>
<thead>
<tr>
<th></th>
<th>6 x 4 TB SSDs/server</th>
<th>24 x 4 TB/ODSA Server</th>
</tr>
</thead>
<tbody>
<tr>
<td>Server cost ($USD, without SSDs)</td>
<td>$10,500</td>
<td>$10,500</td>
</tr>
<tr>
<td>Capacity/Server (TB)</td>
<td>24</td>
<td>96</td>
</tr>
<tr>
<td>#Servers needed for 1 PB SSD</td>
<td>42</td>
<td>11</td>
</tr>
<tr>
<td>Total SSD costs ($USD)</td>
<td>$200,000</td>
<td>$250,000</td>
</tr>
<tr>
<td>Bare server costs ($USD)</td>
<td>$441,000</td>
<td>$115,500</td>
</tr>
<tr>
<td>Total server costs ($USD)</td>
<td>$641,000</td>
<td>$365,500</td>
</tr>
<tr>
<td>Total host maint costs ($USD)</td>
<td>$756,000</td>
<td>$198,000</td>
</tr>
<tr>
<td>Total TCO</td>
<td>$1,397,000</td>
<td>$563,500</td>
</tr>
<tr>
<td>TCO reduction</td>
<td></td>
<td>59.7%</td>
</tr>
</tbody>
</table>

### Assume:
- ODSA accelerator adds 25% to cost of SSD storage
- Performance of 1 ODSA Server is comparable to 4 standard servers.

### Notes:
- Fewer servers leads to simpler networking and additional savings
- Higher performance on ODSA servers would lead to even better ROI
- Even slower performance might still be a better solution
What Will The Future Bring

Chiplet technology will allow for rapid creation of accelerators

We must have an ecosystem of chiplets, corresponding software, and standards that cover many different areas:

- CPU
- GPU
- NPU
- Network
- FPGA
- Other?
A journey shared takes us beyond