The five-minute rule thirty years later

Raja Appuswamy, Renata Borovica-Gajic, Goetz Graefe, and Anastasia Ailamaki
The five-minute rule in 1987

- Storage hardware: Two-tier hierarchy
  - 1MB RAM: $5,000 ~ $5,000/MB
  - 180MB HDD: $30,000 ~ $160/MB

- Optimization problem
  "When does it make sense to cache data in DRAM?"

- Gray & Putzolu’s answer
  "Pages referenced every 5 minutes should be memory resident"
Five-minute rule formulation

Break-even Reference Interval (seconds) =

\[
\frac{\text{PagesPerMBofRAM}}{\text{AccessPerSecondPerDisk}} \times \frac{\text{Technology ratio}}{\text{PricePerDiskDrive}} \times \frac{\text{Economic ratio}}{\text{PricePerMBofDRAM}}
\]
Five-minute rule formulation

Break-even Reference Interval (seconds) = (400 secs)

\[
\text{PagesPerMBofRAM} \div \text{AccessPerSecondPerDisk} \times \text{PricePerDiskDrive} \div \text{PricePerMBofDRAM}
\]

Technology ratio

Economic ratio

Popular rule of thumb for engineering data management systems
Modern storage hierarchy

Mutitier hierarchy with price and performance matching workload requirements
Agenda

• Revisiting the five-minute rule
  – DRAM-HDD break-even interval after 30 years
  – DRAM-SSD, HDD-SSD break-even intervals

• Five-minute rule and the performance tier
  – Break-even intervals with NVDIMM & NVMe SSD

• Five-minute rule and the capacity tier
  – Break-even intervals with Cold Storage, LTO-7 tape
### Storage hardware 30 years later

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Disk (then)</th>
<th>Disk (now)</th>
<th>DRAM (then)</th>
<th>DRAM (now)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit cost ($)</td>
<td>$30,000</td>
<td>$49</td>
<td>$5,000</td>
<td>$80</td>
</tr>
<tr>
<td>Unit capacity</td>
<td>180MB</td>
<td>2TB</td>
<td>1MB</td>
<td>16GB</td>
</tr>
<tr>
<td>Random IO/s</td>
<td>15</td>
<td>200</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

- Capacity: ↑10,000×, Cost: ↓1,000×, HDD Performance: ↑10×
**Five-minute rule 30 years later**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Disk (then)</th>
<th>Disk (now)</th>
<th>DRAM (then)</th>
<th>DRAM (now)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit cost ($)</td>
<td>$30,000</td>
<td>$49</td>
<td>$5,000</td>
<td>$80</td>
</tr>
<tr>
<td>Unit capacity</td>
<td>180MB</td>
<td>2TB</td>
<td>1MB</td>
<td>16GB</td>
</tr>
<tr>
<td>Random IO/s</td>
<td>15</td>
<td>200</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

- Capacity: ↑10,000×, Cost: ↓1,000×, HDD Performance: ↑10×

<table>
<thead>
<tr>
<th>Page size (4KB)</th>
<th>Then</th>
<th>Now</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAM-HDD</td>
<td>5 mins</td>
<td>5 hours</td>
</tr>
</tbody>
</table>

- RAM-HDD break-even 60× higher due to fall in DRAM price

*Store only extremely “cold” data in HDD*
Five-minute rule with SATA SSD

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Disk (now)</th>
<th>DRAM (now)</th>
<th>SATA SSD (now)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit cost ($)</td>
<td>$49</td>
<td>$80</td>
<td>560</td>
</tr>
<tr>
<td>Unit capacity</td>
<td>2TB</td>
<td>16GB</td>
<td>800GB</td>
</tr>
<tr>
<td>Cost/MB</td>
<td>0.00002</td>
<td>0.005</td>
<td>0.0007</td>
</tr>
<tr>
<td>Random IO/s</td>
<td>200</td>
<td>-</td>
<td>67k/20k</td>
</tr>
</tbody>
</table>

• Two properties of SSDs
  • Middleground between DRAM and HDD w.r.t cost/MB
  • 100-1000× higher random IOPS than HDD
Five-minute rule with SATA SSD

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Disk (now)</th>
<th>DRAM (now)</th>
<th>SATA SSD (now)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit cost ($)</td>
<td>$49</td>
<td>$80</td>
<td>560</td>
</tr>
<tr>
<td>Unit capacity</td>
<td>2TB</td>
<td>16GB</td>
<td>800GB</td>
</tr>
<tr>
<td>Cost/MB</td>
<td>0.00002</td>
<td>0.005</td>
<td>0.0007</td>
</tr>
<tr>
<td>Random IO/s</td>
<td>200</td>
<td>-</td>
<td>67k/20k</td>
</tr>
</tbody>
</table>

- Two properties of SSDs
  - Middleground between DRAM and HDD w.r.t cost/MB
  - 100-1000× higher random IOPS than HDD

- Two new rules with SSDs
  - DRAM-SSD rule: SSD as a primary store
  - SSD-HDD rule: SSD as a cache
# Break-even interval for SATA SSD

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Disk (now)</th>
<th>DRAM (now)</th>
<th>SATA SSD (now)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit cost ($)</td>
<td>$49</td>
<td>$80</td>
<td>560</td>
</tr>
<tr>
<td>Unit capacity</td>
<td>2TB</td>
<td>16GB</td>
<td>800GB</td>
</tr>
<tr>
<td>Cost/MB</td>
<td>0.000002</td>
<td>0.005</td>
<td>0.00007</td>
</tr>
<tr>
<td>Random IO/s</td>
<td>200</td>
<td>-</td>
<td>67k (r)/20k (w)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Page size (4KB)</th>
<th>Then</th>
<th>Now</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAM-HDD</td>
<td>5 mins</td>
<td>5 hours</td>
</tr>
<tr>
<td>RAM-SSD</td>
<td>-</td>
<td>7 m (r)/24m (w)</td>
</tr>
</tbody>
</table>

5-minute rule now ~applicable to SATA SSD
Break-even interval for SATA SSD

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Disk (now)</th>
<th>DRAM (now)</th>
<th>SATA SSD (now)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit cost ($)</td>
<td>$49</td>
<td>$80</td>
<td>560</td>
</tr>
<tr>
<td>Unit capacity</td>
<td>2TB</td>
<td>16GB</td>
<td>800GB</td>
</tr>
<tr>
<td>Cost/MB</td>
<td>0.00002</td>
<td>0.005</td>
<td>0.00007</td>
</tr>
<tr>
<td>Random IO/s</td>
<td>200</td>
<td>-</td>
<td>67k (r)/20k (w)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Page size (4KB)</th>
<th>Then</th>
<th>Now</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAM-HDD</td>
<td>5 mins</td>
<td>5 hours</td>
</tr>
<tr>
<td>RAM-SSD</td>
<td>-</td>
<td>7 m (r)/24m (w)</td>
</tr>
<tr>
<td>SSD-HDD</td>
<td>-</td>
<td>1 day</td>
</tr>
</tbody>
</table>

5-minute rule now ~applicable to SATA SSD
With 1 day interval, all active data will be in RAM/SSD
Agenda

• Revisiting the five-minute rule
  – DRAM-HDD break-even interval after 30 years
  – DRAM-SSD, HDD-SSD break-even intervals

• Five-minute rule and the performance tier
  – Break-even intervals with NVDIMM & NVMe SSD

• Five-minute rule and the capacity tier
  – Break-even intervals with Cold Storage, LTO-7 tape
Trends in performance tier

• SSDs inching closer to the CPU
  – SATA -> SAS/FiberChannel -> PCIe -> NVMe -> DIMM
  – NVMe PCIe SSDs are server accelerators of choice

<table>
<thead>
<tr>
<th>Device</th>
<th>Capacity</th>
<th>Price ($)</th>
<th>IOPS (k) r/w</th>
<th>B/W (GBps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SATA SSD</td>
<td>800GB</td>
<td>560</td>
<td>67/20</td>
<td>500/460</td>
</tr>
<tr>
<td>Intel 750</td>
<td>1TB</td>
<td>630</td>
<td>460/290</td>
<td>2.5/1.2</td>
</tr>
</tbody>
</table>
Trends in performance tier

• SSDs inching closer to the CPU
  – SATA -> SAS/FiberChannel -> PCIe -> NVMe -> DIMM
  – NVMe PCIe SSDs are server accelerators of choice

• Storage Class Memory devices (ex: 3D Xpoint)
  – Faster than Flash, Denser than DRAM, and non-volatile
  – Standardized, byte-addressable, NVDIMM-P soon

<table>
<thead>
<tr>
<th>Device</th>
<th>Capacity</th>
<th>Price ($)</th>
<th>IOPS (k)</th>
<th>B/W (GBps)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>r/w</td>
<td></td>
</tr>
<tr>
<td>SATA SSD</td>
<td>800GB</td>
<td>560</td>
<td>67/20</td>
<td>500/460</td>
</tr>
<tr>
<td>Intel 750</td>
<td>1TB</td>
<td>630</td>
<td>460/290</td>
<td>2.5/1.2</td>
</tr>
<tr>
<td>Intel P4800X</td>
<td>384GB</td>
<td>1520</td>
<td>550/500</td>
<td>2.5/2</td>
</tr>
</tbody>
</table>
## Break even interval for PCIe SSD/NVM

<table>
<thead>
<tr>
<th>Device</th>
<th>Capacity</th>
<th>Price ($)</th>
<th>IOPS (k) r/w</th>
<th>B/W (GBps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SATA SSD</td>
<td>800GB</td>
<td>560</td>
<td>67/20</td>
<td>500/460</td>
</tr>
<tr>
<td>Intel 750</td>
<td>1TB</td>
<td>630</td>
<td>460/290</td>
<td>2.5/1.2</td>
</tr>
<tr>
<td>Intel P4800X</td>
<td>384GB</td>
<td>1520</td>
<td>550/500</td>
<td>2.5/2</td>
</tr>
</tbody>
</table>

### Page size (4KB)

<table>
<thead>
<tr>
<th>RAM-SATA SSD</th>
<th>7 m (r) / 24m (w)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAM-Intel 750</td>
<td>41 s (r) / 1m (w)</td>
</tr>
<tr>
<td>RAM-P4800X</td>
<td>47 s (r) / 52s (w)</td>
</tr>
</tbody>
</table>

DRAM-NVM break-even interval is shrinking
Interval disparity between reads and writes is shrinking
Break even interval for PCIe SSD/NVM

<table>
<thead>
<tr>
<th>Device</th>
<th>Capacity</th>
<th>Price ($)</th>
<th>IOPS (k) r/w</th>
<th>B/W (GBps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SATA SSD</td>
<td>800GB</td>
<td>560</td>
<td>67/20</td>
<td>500/460</td>
</tr>
<tr>
<td>Intel 750</td>
<td>1TB</td>
<td>630</td>
<td>460/290</td>
<td>2.5/1.2</td>
</tr>
<tr>
<td>Intel P4800X</td>
<td>384GB</td>
<td>1520</td>
<td>550/500</td>
<td>2.5/2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Page size (4KB)</th>
<th>Now</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAM-SATA SSD</td>
<td>7 m (r) / 24m (w)</td>
</tr>
<tr>
<td>RAM-Intel 750</td>
<td>41 s (r) / 1m (w)</td>
</tr>
<tr>
<td>RAM-P4800X</td>
<td>47 s (r) / 52s (w)</td>
</tr>
</tbody>
</table>

DRAM-NVM break-even interval is shrinking.
Interval disparity between reads and writes is shrinking.

*Impending shift from DRAM to NVM-based data management engines*
Agenda

• Revisiting the five-minute rule
  – DRAM-HDD break-even interval after 30 years
  – DRAM-SSD, HDD-SSD break-even intervals

• Five-minute rule and the performance tier
  – Break-even intervals with NVDIMM & NVMe SSD

• Five-minute rule and the capacity tier
  – Break-even intervals with Cold Storage, LTO-7 tape
Trends in high-density storage

• HDD scaling falls behind Kryder’s rate
  – PMR provides 16% improvement in areal density, not 40%
Trends in high-density storage

• HDD scaling falls behind Kryder’s rate
  – PMR provides 16% improvement in areal density, not 40%

• Tape density continues 33% growth rate
  – IBM’s new record: 123 Billion bits/sq. inch
  – But high access latency
Trends in high-density storage

- HDD scaling falls behind Kryder’s rate
  - PMR provides 16% improvement in areal density, not 40%

- Tape density continues 33% growth rate
  - IBM’s new record: 123 Billion bits/sq. inch
  - But high access latency

- Flash density outpacing rest
  - 40% density growth due to volumetric + areal techniques
  - But high cost/GB
Trends in high-density storage

• HDD scaling falls behind Kryder’s rate
  – PMR provides 16% improvement in areal density, not 40%

• Tape density continues 33% growth rate
  – IBM’s new record: 123 Billion bits/sq. inch
  – But high access latency

• Flash density outpacing rest
  – 40% density growth due to volumetric + areal techniques
  – But high cost/GB

• Cold storage devices (CSD) filling the gap
  – 1,000 high-density SMR disks in MAID setup
  – PB density, 10s latency, 2-10GB/s bandwidth
# Break-even interval for tape

<table>
<thead>
<tr>
<th>Metric</th>
<th>DRAM</th>
<th>HDD</th>
<th>SpectraLogic T50e tape library</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit capacity</td>
<td>16GB</td>
<td>2TB</td>
<td>10 * 15TB</td>
</tr>
<tr>
<td>Unit cost ($)</td>
<td>80</td>
<td>50</td>
<td>11,000</td>
</tr>
<tr>
<td>Latency</td>
<td>100ns</td>
<td>5ms</td>
<td>65s</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>100GB/s</td>
<td>200MB/s</td>
<td>4 * 750 MB/s</td>
</tr>
</tbody>
</table>

- DRAM-tape break-even interval: 300 years!
  
  "Tape: The motel where data checks in and never checks out"
  
  - Jim Gray

- Kaps is not the right metric for tape
  
  - Maps, TB-scan better
## Alternate comparison metrics

<table>
<thead>
<tr>
<th>Metric</th>
<th>DRAM</th>
<th>HDD</th>
<th>SpectraLogic T50e tape library</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit capacity</td>
<td>16GB</td>
<td>2TB</td>
<td>10 * 15TB</td>
</tr>
<tr>
<td>Unit cost ($)</td>
<td>80</td>
<td>50</td>
<td>11,000</td>
</tr>
<tr>
<td>Latency</td>
<td>100ns</td>
<td>5ms</td>
<td>65s</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>100GB/s</td>
<td>200MB/s</td>
<td>4 * 750 MB/s</td>
</tr>
<tr>
<td>$/Kaps (amortized)</td>
<td>9e-14</td>
<td>5e-9</td>
<td>8e-3</td>
</tr>
<tr>
<td>$/TBScan (amortized)</td>
<td>8e-6</td>
<td>3e-3</td>
<td>3e-2</td>
</tr>
</tbody>
</table>

**HDD 1,000,000× cheaper w.r.t Kaps, only 10× w.r.t TBScan**

**HDD—tape gap shrinking for sequential workloads**

Implications for the capacity tier

• Traditional tiering hierarchy
  – HDD based capacity tier. Tape, CSD only used in archival.
Implications for the capacity tier

• Traditional tiering hierarchy
  – HDD based capacity tier. Tape, CSD only used in archival.

• Clear division in workloads
  – Only non-latency sensitive, batch analytics in capacity tier
Implications for the capacity tier

• Traditional tiering hierarchy
  – HDD based capacity tier. Tape, CSD only used in archival.

• Clear division in workloads
  – Only non-latency sensitive, batch analytics in capacity tier

• Is it economical to merge the two tiers?
  – “40% cost savings by using a cold storage tier” [Skipper, VLDB’16]
Implications for the capacity tier

• Traditional tiering hierarchy
  – HDD based capacity tier. Tape, CSD only used in archival.

• Clear division in workloads
  – Only non-latency sensitive, batch analytics in capacity tier

• Is it economical to merge the two tiers?
  – “40% cost savings by using a cold storage tier” [Skipper, VLDB’16]

• Can batch analytics be done on tape/CSD?
  – Query Execution in Tertiary Memory Databases [VLDB’96]
  – Skipper: Cheap data analytics over cold storage devices [VLDB’16]
  – Nakshatra: Running batch analytics on an archive [MASCOTS’14]

Time to revisit traditional capacity—archival division of labor
Summary

• Growing DRAM-HDD & shrinking DRAM-NVM intervals
  
  **Most performance critical data will sit in SSD/NVM**

• Rapid improvements in SSD/NVM density
  
  **All randomly accessed data can sit in SSD/NVM**

• Shrinking HDD—tape/CSD difference w.r.t $/TBscan
  
  **Can merge archival+capacity tier into cold storage tier**

  **Sequential batch analytics can be hosted on new tier**

Five-minute rule suggests impending consolidation in the storage hierarchy