SSDs EXCEL IN BIG DATA ARCHITECTURE

Boosting cost/performance with SSDs is all about **LOCATION, LOCATION, LOCATION**

Here are 8 critical spots in big data architecture....

1 SCALE UP

One larger multicore, all-SSD node in a Hadoop cluster can take on bigger jobs or chaining, freeing up smaller nodes for easier jobs [1]

2 VIRTUALIZATION

SSDs mitigate ill effects of smaller I/O blocks and larger number of requests in Xen [2]

3 HYBRID DATABASE

Keeping smaller files and indices in RAM, with larger files stored on SSDs, speeds up NoSQL [8]

4 SPILLING RDDS

Spark running on Amazon EC2 nodes backed by Amazon S3 all-SSD storage sets records [7]

5 MAP REDUCE

Targeting intermediate shuffle results produced by MapReduce algorithms at an SSD reduces severe loads on overwhelmed HDDs [3]

6 TIERED STORAGE

Aiming more HDFS requests at PCIe SSDs in a hybrid cluster increases average I/O rate [4]

7 SMALL "HOT" FILES

Caching Facebook messages on an SSD triples HBase performance through reduced latency [5]

8 TIGHTLY COUPLED

MIT tests FPGA fabric and local ARM cores on SSDs to pre-process data and speed searches [6]

BIG DATA CLUSTERS WITH SSDS IN CRITICAL LOCATIONS COMPLETE MORE REQUESTS,

SSDs with V-NAND unlock big data performance with:

- ↑ HIGHER IOPS
- ↓ LOWER LATENCY
- \+ MORE DWPD

For more details on findings from these research studies **samsung.com/enterprisessd**

READ THE WHITE PAPER TO LEARN MORE

1 <Scale Up>
"Scale-up vs Scale-out for Hadoop: Time to rethink?", Appuswamy et al, Microsoft Research, presented at SoCC '13, October 2013, http://www.msr-waypoint.com/pubs/204499/a20-appuswamy.pdf

2 <Virtualization>
"Performance Implications of SSDs in Virtualized Hadoop Clusters", Ahn et al, Samsung Electronics, presented at IEEE BigData Congress, June 2014, http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=6906832

3 <MapReduce>
"The Truth About MapReduce Performance on SSDs", Karthik Kambatla and Yanpei Chen, Cloudera Inc. and Purdue University, presented at LISA 14 sponsored by USENIX, November 2014, https://www.usenix.org/conference/lisa14/conference-program/presentation/kambatla

4 <Tiered Storage>
"hatS: A Heterogeneity-Aware Tiered Storage for Hadoop", Krish et al, Virginia Tech, presented at IEEE/ACM CCGrid2014, May 2014, http://people.cs.vt.edu/butta/docs/ccgrid2014-hats.pdf

5 <Small, "Hot" Files>
"Analysis of HDFS under HBase: A Facebook Messages Case Study", Harter et al, Facebook Inc. and University of Wisconsin, Madison, presented at FAST '14 sponsored by USENIX, February 2014, http://research.cs.wisc.edu/adsl/Publications/fbmessages-fast14.pdf

6 <Tightly Coupled>
"Cutting cost and power consumption for big data", Larry Hardesty, Massachusetts Institute of Technology, July 10, 2015, http://news.mit.edu/2015/cutting-cost-power-big-data-0710

7 <Spilling RDDs>
"Spark the fastest open source engine for sorting a petabyte", Databricks, November 2014, https://databricks.com/blog/2014/10/10/spark-petabyte-sort.html

8 <Hybrid Database>
Aerospike web site, http://www.aerospike.com/flash-optimized/

SAMSUNG