

A Case for Flash Memory SSD in OLTP Applications

One FlashSSD can beat Ten 15K RPM HDDs - Performance, Price, Power -

Sang-Won Lee

http://icc.skku.ac.kr/~swlee

(Joint Work: Bongki Moon@Univ. of Arizona, Tony Park@Indilinx)







Santa Clara, CA USA August 2009

Flash Memory Markets

- Mobile devices
- PC, Laptop
- Enterprise server storages

A Case for Flash Memory SSDs in OLTP Applications 2





A Gloomy Marketing Strategy

- One SSD vs. One Harddisk
 - e.g One SSD's Capacity: 32 GB, 80 GB
- Someone says that "SSD can penetrate the market only when it matches HDD price"
 - Partially true in PC / Laptop market
- Under this strategy, the market would be invulnerable to SSD



Three Truths / Myths on SSD

- SSD is expensive!
- SSD's power consumption is non-trivial!
- Write performance is problematic!



Motivations

- "FlashSSD's message is still unclear in the market"
 - [Personal Communication] Ken Salem, University of Waterloo
- It is urgent to develop "the case for flash memory SSD" (or killer applications) and "the right message"
- Debunk those myths on SSDs
 - From OLTP Perspectives



IBM for TPC-C (2008 Dec.)

- 800 TB Storage
 - 11K 73.4GB disks (each 15k rpm)
- Total cost: 35M \$
 - Server HW: 12M \$
 - Server SW: 2M \$
 - Storage: 20M \$
 - Client HW/SW: 1M \$
 - To improve IOPS
- They buy IOPS, not capacity!

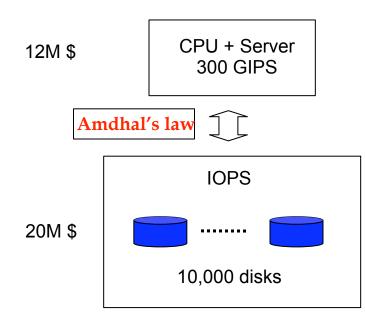
TUM		IBM Power 595			TPC-C Rev. 5.9		
	N	Model 9119-FHA		Report Date: June 10, 2008			
Total System Cast	TPC-C The	roughput	Price/Perfo	rmance	Availability Date		
\$17,111,788 USD	6,085,	166	\$2.81 USD		December 10, 2000		
Database Server Processor Chip/Coro/The		abase	Operating System	Other	Software	No. User	
32/64/128 POWER6 5.0GHz	DB	2.9.5	AIX 5L V5.3	Microsoft Visual C++ Microsoft COM+ 5,184		5,184,000	
		_		_	_		
SEE Common Bion Structures and Charl court of the La Court Court own of the La Court Court own of the Court Court of the Court Part of the Court Part of the Court Integrated data port the team	a Kaon'ne	30 Pec 64 5.000 3040 40 8 1.4000 61 400-046 p	Reset 201 Server consol Oppe with a Powerse Conset and Another before thermal of Power Conset of Power Conset of Power Conset of Power Conset 1000 Ethermed Adaptives	754 KNI Sunlare	Booage ten Borage chance Strage (Saloo Lo) 154 PPM alloge (m	Peto	
Bin Species 20 Out on 2 2004 that 488 L 2 0 sets 1028 Memory 73 486 545 0n Integrated due por to 1001	a Kaon'ne	30 Pec 64 5.000 3040 40 8 1.4000 61 400-046 p	Consol Oppo wills a Control per chip solicit-terret Reference Consol Reference Acceptors 1000 Ethermed Acceptors 1000 Ethermed Acceptors	Plai title Sphere 10,002 75400	Stringe Disation Elit	NET	
Ella faciliaria (6) Danii-cora 2 dilarte intent ettali L2 Caste 1020 Matterny 75,4000 SAS (2)	a Kaon'ne	20 Pic 64 5 100 5000 81 420 0as p 10 10 100 50 10 100 50 10 100	Consol Oppo wills a Control per chip solicit-terret Reference Consol Reference Acceptors 1000 Ethermed Acceptors 1000 Ethermed Acceptors	Plai title Sphere 10,002 75400	Strage (6400 D)	Tients	
System	to saon"**	20 PK 43 500 40 81400 40 420-040 10 10100 50 10100 50 10100 10 10100 10 10100 10 10100	Consol Oppo wills e Montellin Serie 13 Garba per chip social Mannus Interna Scola Consol Interna Scola Consol 1000 Ethermet Adapters 1000 Ethermet Adapters	Figure 75400	therape (Selice For the Arms allope (m	Tients ption B L2 Xeon	
System Components Processors System Components Processors Chips	Quantity	20 PK 43 500 40 81400 40 420-040 10 10100 50 10100 50 10100 10 10100 10 10100 10 10100	Some Organ will: I Some Der Ang I Some Der Ang Sone Some Some Some Some Some Some Some Some Some Some Some Provider Some So	Eact Quantity	h of the 128 C Descrit 2.0GHz 4M	Tients ption B L2 Xeon ssor	
System Components System Components Processors Chips /Cores/Threads	Quantity 32/64/128	Server States a table a table	Construction of the output of	Eacl Quantity 1/2/2	h of the 128 C Descri 2.0GHz 4M Proce	Tients ption B L2 Xeon stor MB	
System Components System Components Processors Chips /Cores/Threads	Quantity 3264/128 64 1 68	Server a tool a tool	A Controller will reference of the second reference of the second of the second of the second of the second of the second of	Each Quantity 1/2/2 2	h of the 128 C Descri 2.0Ek 4M Proce 512 1	Tients ption B L2 Xeon mor MB arroller	
System Components Processors Chips /Cores/Threads Memory Disk Controllers	Quantity 32864/128 64 1 68 68 8	2014 64 500 2014 21 4000 21 4000 21000 21000 21000 21000 21000 21000 210000 2100000000	err Honoral Constant Honoral Constant Honoral Constant Honoral Constant Honoral Constant Honoral Constant Honor Constant Honor Constant Honoral Constant Honoral Constant Honoral Constant Honoral Ho	Each Quantity 1/2/2 2 1	h of the 128 C Descri 2.0GHz 4M Proce 5121 SAS Co	Thents ption B L2 Xeon mor MB arrollor RPM SAS	







 For balanced systems, OLTP systems pay huge \$\$\$ on disks for high IOPS; IOPS crisis would be worse and worse

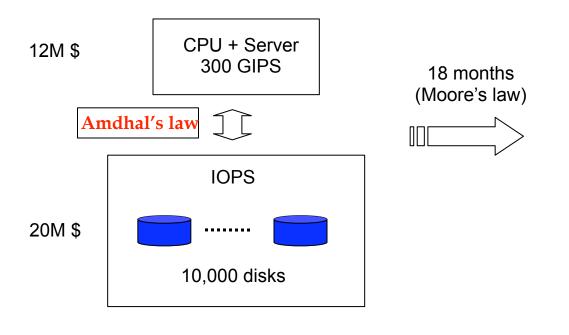


A balanced state

A Case for Flash Memory SSDs in OLTP Applications **7**



 For balanced systems, OLTP systems pay huge \$\$\$ on disks for high IOPS; IOPS crisis would be worse and worse

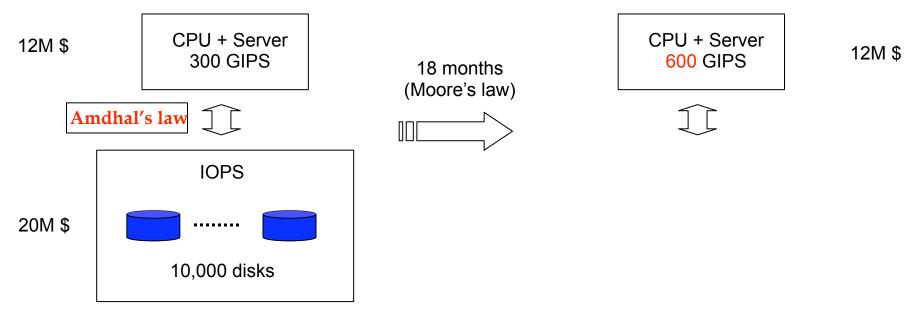


A balanced state

A Case for Flash Memory SSDs in OLTP Applications **7**



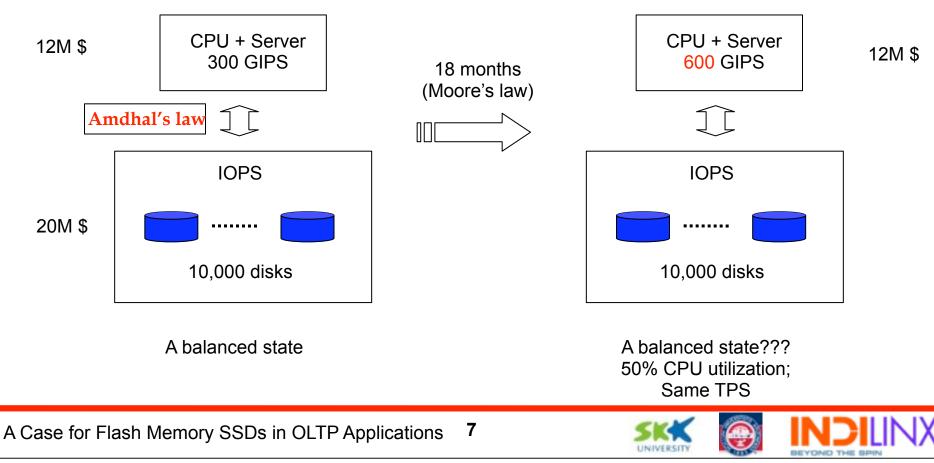
 For balanced systems, OLTP systems pay huge \$\$\$ on disks for high IOPS; IOPS crisis would be worse and worse



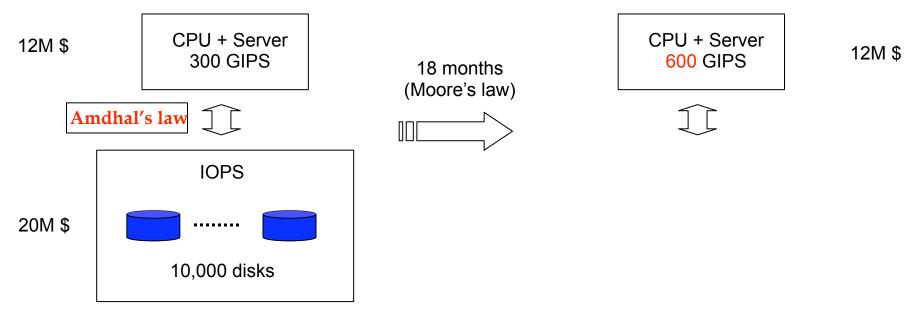




 For balanced systems, OLTP systems pay huge \$\$\$ on disks for high IOPS; IOPS crisis would be worse and worse



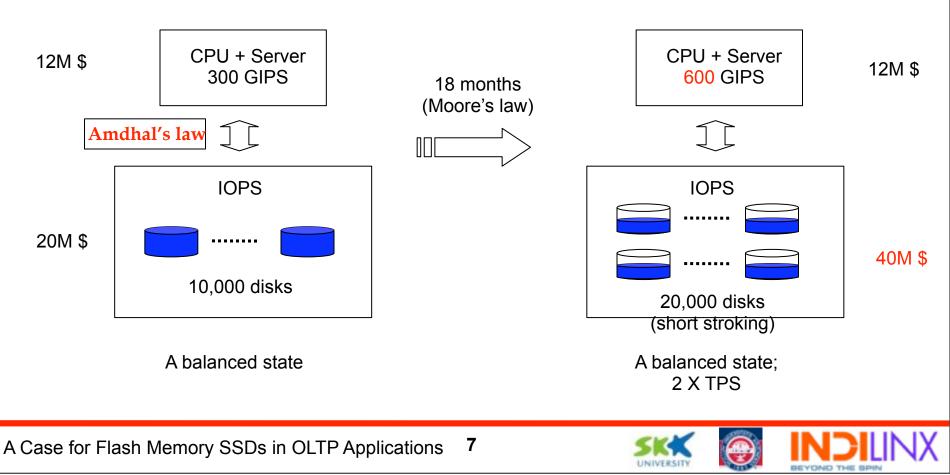
 For balanced systems, OLTP systems pay huge \$\$\$ on disks for high IOPS; IOPS crisis would be worse and worse







 For balanced systems, OLTP systems pay huge \$\$\$ on disks for high IOPS; IOPS crisis would be worse and worse



Indilinx SSD vs. HDD

• INDILINX Barefoot Controller: 2.5" 32GB SLC SSD with



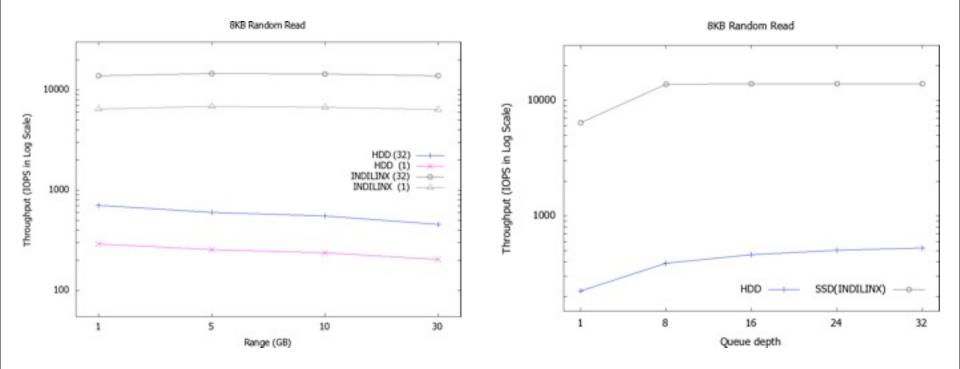
 HDD: 15K rpm Seagate 73.4GB SAS Cheetah 15K.5 model (model no.: ST373455SS)

A Case for Flash Memory SSDs in OLTP Applications 8



Simple IOPS(1)

- Random read vs. data range vs. NCQ queue depth
 - 20 ~ 30 times faster (8KB)

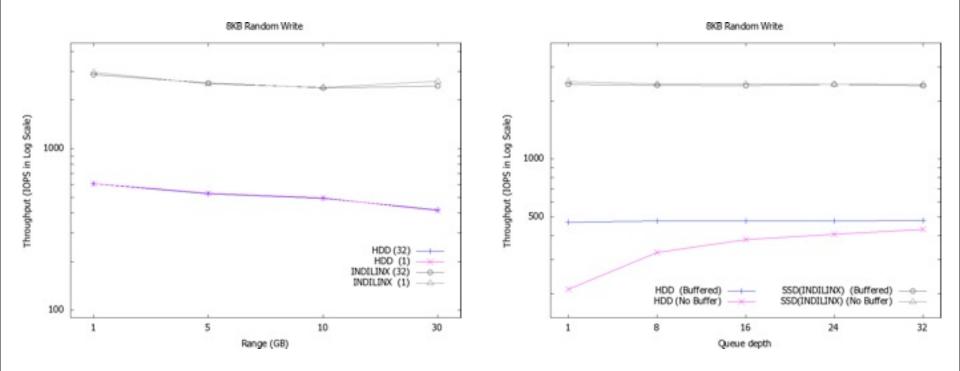


A Case for Flash Memory SSDs in OLTP Applications 9



Simple IOPS(2)

- Random write vs. data range vs. NCQ queue depth
 - $5 \sim 6$ times faster (8KB)



A Case for Flash Memory SSDs in OLTP Applications 10



TPC-C Bechmark

- TPC = Transaction Processing Performance Council
 - TPC-C: De facto industry standard benchmark for OLTP performance
 - 5 types of transactions:
 - ✓ Read only: Order-status(4%), Stock-level(4%)
 - Read/Write mixed: New-order(45%, heavy write),
 Payment(43%, light write), Delivery(4%, medium write)
- IO Characteristics
 - Unit of IO: 2 ~ 8K page
 - Ratio of read and write ~~ 1:1



TPC-C Benchmark: Experimental Setups

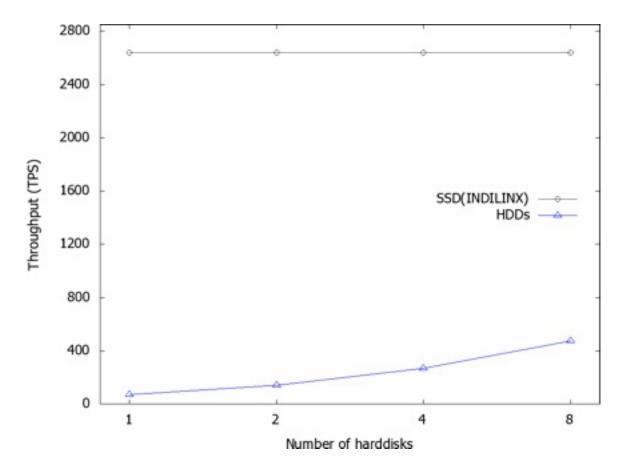
- CPU: Intel Core 2 Quad Q6600
- Mother Board: ASUS P5K-E
- RAM: Samsung DDR2 1GB × 2 (2GB)
- OS: Oracle Enterprise Linux 5.1
- DBMS: Oracle 10g R2 (10.2.0.1.0) for Linux x86
- RAID Controller: Intel RAID Controller SRCSASRB
- TPC-C benchmark software: BMFactory
 - 10GB database
 - 100MB buffer
- 8-HDDs vs. 1-SSD





TPC-C Performance: Read Only TPS

Order_Status: One SSD vs.8 HDDs = 5:1

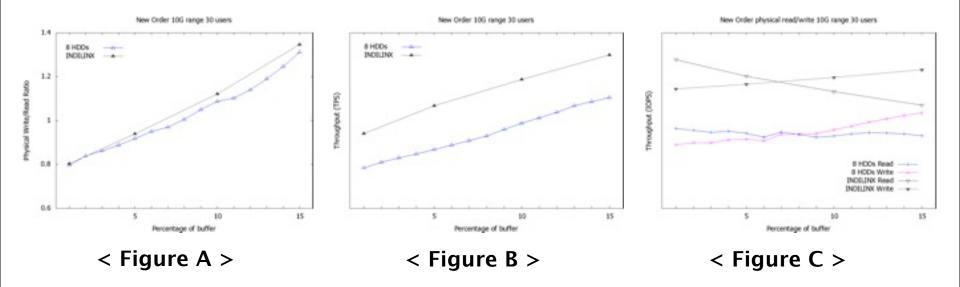


A Case for Flash Memory SSDs in OLTP Applications 13



TPC-C Performance: Read/Write TPS

- New_Order: One SSD vs. 8 HDDs = 1.5 ~ 2 : 1
 - Figure A: large buffer means higher physical W/R ratio
 - Figure B: TPS increases
 - Figure C: But, the performance improvement ratio in SSD lags that of HDD because of random write bottleneck

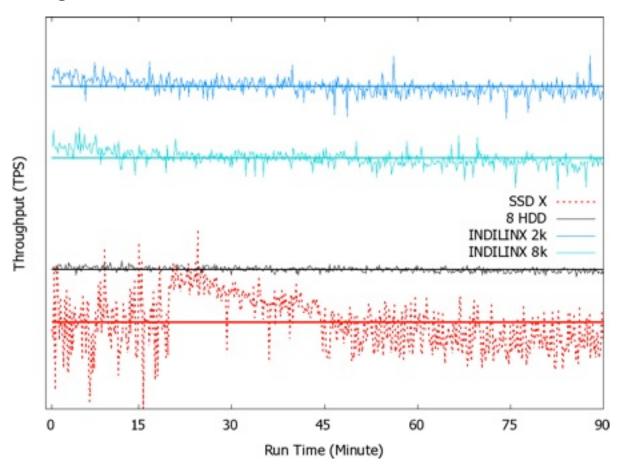


A Case for Flash Memory SSDs in OLTP Applications 14



TPC-C Performance: Read/Write TPS

TPS change over time



A Case for Flash Memory SSDs in OLTP Applications 15



Power Consumption in OLTP

- Meikel Poess et al., Energy Cost: The Key Challenge of Today's Data Centers: A Power Consumption Analysis of TPC-C Results, VLDB 2008
 - In OLTP, storage component consumes 80% of the whole OLTP system
 - Energy metrics will be added in future TPC-C benchmark
- One SSD vs. 8 HDDs
 - Performance: SSD >> 8 HDDs
 - Power: <u>SSD(5W</u>) << HDD(104W)</p>
- In other words, SSD shows
 - very low (Watt or \$) / IO / sec



Conclusion

- OLTP: The Case for Flash Memory SSD
- One FlashSSD can beat 10 15K RPM harddisks
 - Performance, prices, and, power consumption
- The key metric in OLTP storage
 - IOPS / GB, rather than sequential bandwidth and capacity
- We should be more enthusiastic in developing the various cases, where SSD is definitely winner both in terms of price and performances

