



IntelliMagic

Storage Intelligence



# Solid Planning for Solid State Disks

Dr. Gilbert Houtekamer and Wim Oudshoorn  
IntelliMagic BV, The Netherlands

GSE Keep IT Going, 29 oktober 2009

© IntelliMagic 2009



# Objectives

- Introduce Solid State Technology / Enterprise Flash Drives
- Discuss impact on enterprise storage configuration design
- How to justify SSD / EFD to yourself and to your manager
- Migrate Files/Data sets, LUN/volumes or storage groups?
- Sample migration study with actual z/OS data from large financial institution.



# Technology



# High Expectations

Intel co-founder Gordon Moore said:

- “There have been a few times in the history of computing when a new technology becomes completely pivotal to changing the PC platform and the user experience. Solid State Drives have this capability”

A lot of development money goes into SSD, e.g.

- SAN JOSE and SANTA CLARA, Calif. -- Dec. 2, 2008 - Intel Corporation and Hitachi Global Storage Technologies (Hitachi GST) today announced plans to jointly develop and deliver Serial Attached SCSI (SAS) and Fibre Channel (FC) enterprise-class solid-state drives (SSDs) for servers, workstations and storage systems.



## Why is SSD so important?

- Tremendous performance benefits, especially for random I/O
- Can provide very high (read) data rates through parallel reading of cells (really exploit new SAS/FC/SATA speeds)
- No moving parts, potential for much lower manufacturing costs
- In our view, it is likely to displace all high-end FC/SAS drives within 5 years

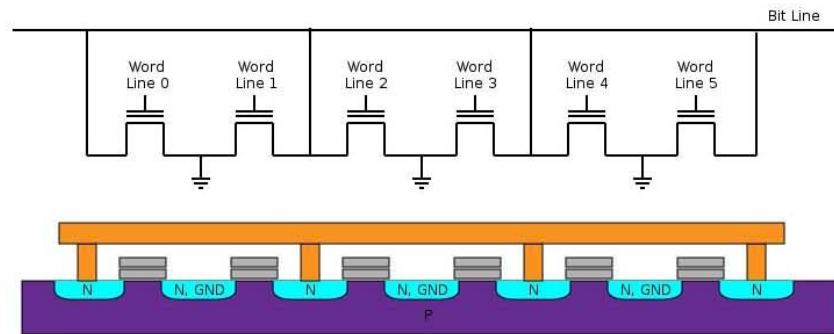
But current disk subsystems are designed for HDD, not SSD



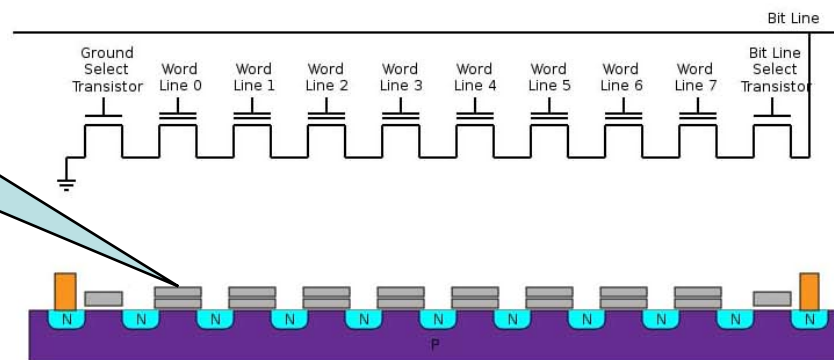
## What is it?

- A **Solid State Drive** is not a disk, but **Flash memory** packaged with a **disk controller**
- **Flash memory** is non-volatile computer memory that can be electrically erased and reprogrammed
- Flash memory contains large number of **cells**, each cell can be charged (written), and this charge can be detected (read)
- Invented around 1980 by Dr Fujio Masuoka from **Toshiba**, first flash product around 1987

# Cells: NAND and NOR

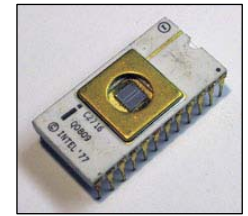


NOR : byte addressable, need large chip area



NAND : block addressable, compact and less expensive

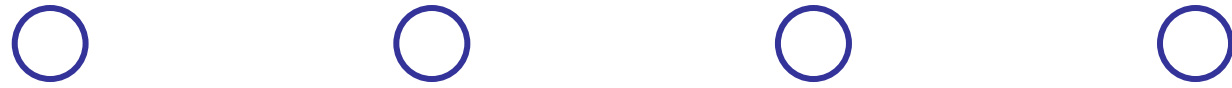
- Flash is derived of byte addressable (E)PROM



NOR technology is still byte addressable

NAND technology is block addressable, like a disk drive

This is where the charge goes



## SLC or MLC

- A Single Level Cell (SLC) is either charged or not
- A Multi Level Cell (MLC) also detects how much charge is present; it can store at least 2 bits
- Which one do you want?
  - SLC is more expensive and supports 100 K R/W cycle
  - MLC is less expensive and supports 10K R/W cycles
  - So an SLC in your mainframe, an MLC in your digital camera
- MLC will become more common as reliability increases
  - Cost argument is too compelling to ignore



- - 
  - 
  -
- ## Erase-Write cycles in Flash
- Flash devices have a limited number of write-erase cycles, typically around 100,000
    - Results from the 'stress' that an erase operation causes
  - To erase the information a rather high voltage is required, which generates heat-stress and small defects in the chip
    - As result of the defects it will become more difficult and at some point impossible to read the information stored
  - Better 'isolation' increases life expectancy and reduces density
    - There are manufacturing choices!
    - Denser chips will have shorter life
    - MLC chips require more accurate 'reading' and hence fail sooner
  - Technology likely to improve a lot in coming years.



## Wear Leveling

- The SSD and/or Flash device firmware employs “*wear leveling*” to ensure all cells are written about the same number of times
- Error correction and redundant cells are used to mask these issues from the user of the Flash / SSD device
- More expensive SSDs have more redundant cells
  - 256 GByte Flash memory sold as 146 Gbyte or 200 GByte SSD



IntelliMagic

## Zeus<sup>IOPS</sup> & Intel<sup>®</sup> Solid State Drives

	Zeus <sup>IOPS</sup> (EMC, IBM, HDS)	X25-E Extreme SATA	X18-M/X25-M SATA (34 nm)
Technology	SLC	SLC	MLC
Capacity	Up to 800 GByte	Up to 64 Gbyte	160 Gbyte
Power (active)	8.4 Watt 5.4 Watt idle	2.4 Watt 0.06 W idle	0.15 Watt 0.075 W idle
Max Mbytes read/write	220 / 110 Gen 1 350 / 300 Gen 2	250 / 170	Up to 250 / 70
Max IOPs	46 K / 16K Gen 1 80K / 40K Gen 2	35K / 3.5K	35K / up to 8.6K
MTBF	Not provided	2 million hours	1.2 million hours
Cost	\$\$\$	\$\$	\$

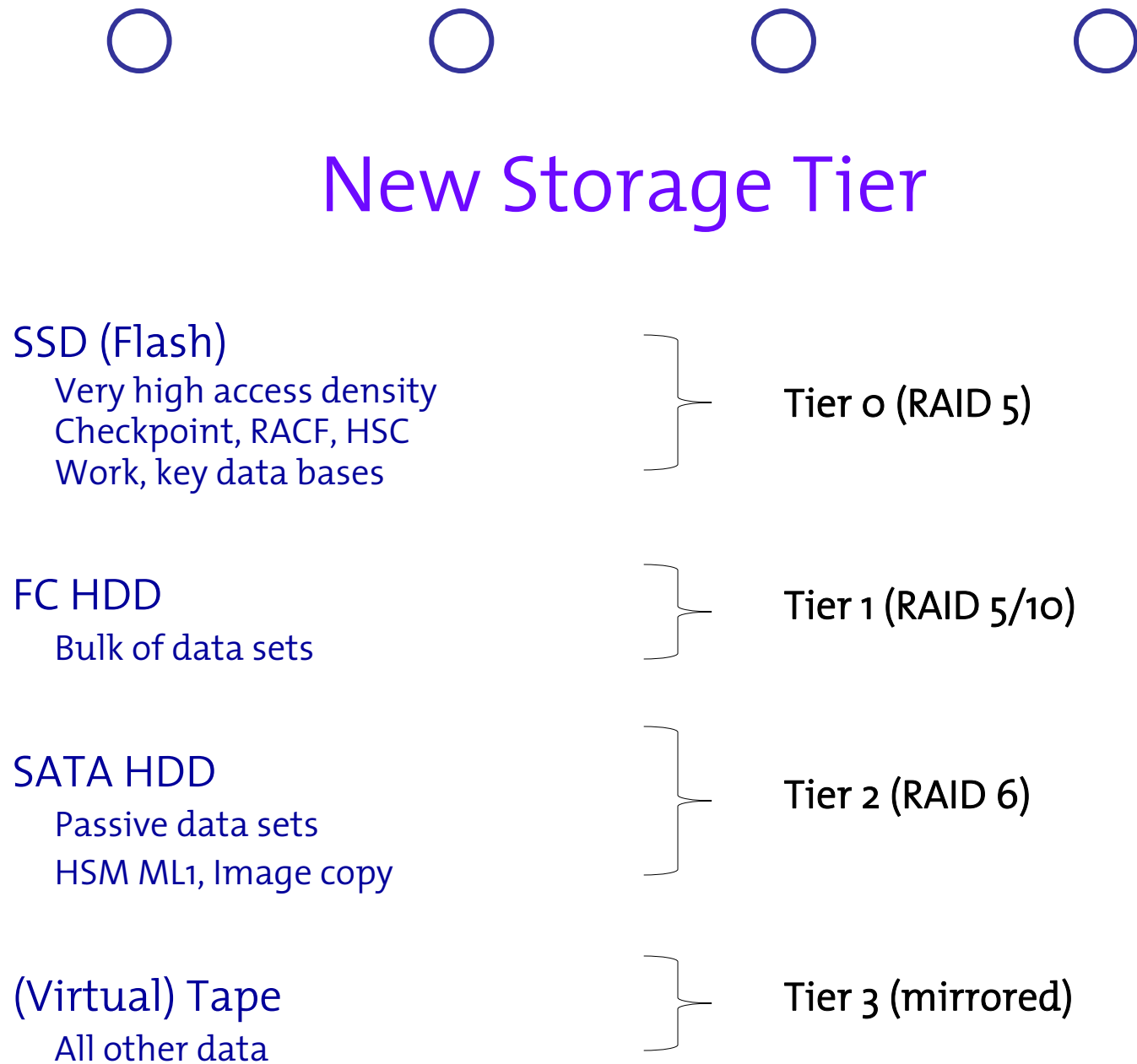


# Configuration design with SSD/EFD



IntelliMagic

Storage Intelligence





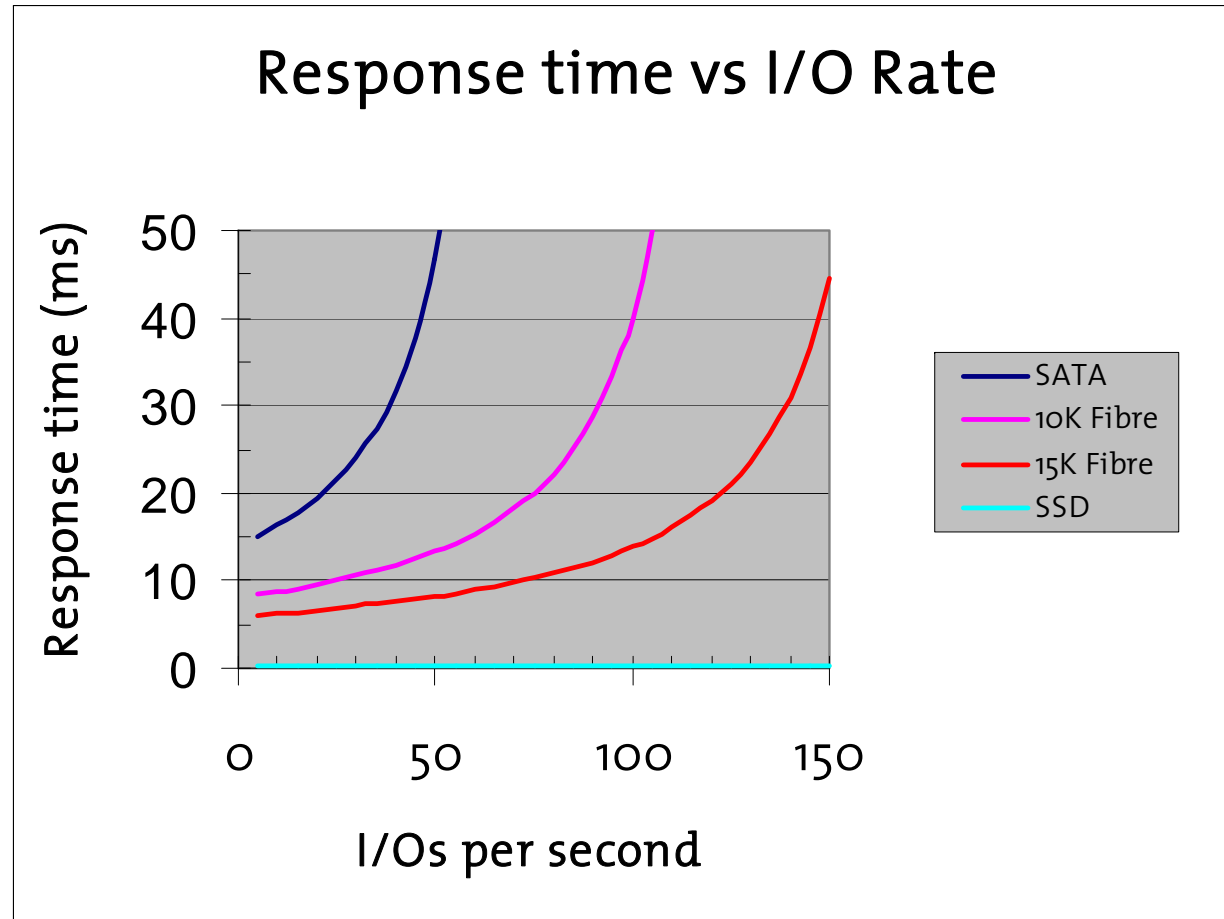
IntelliMagic

Storage Intelligence



# Response Time

Model Results for very small blocks





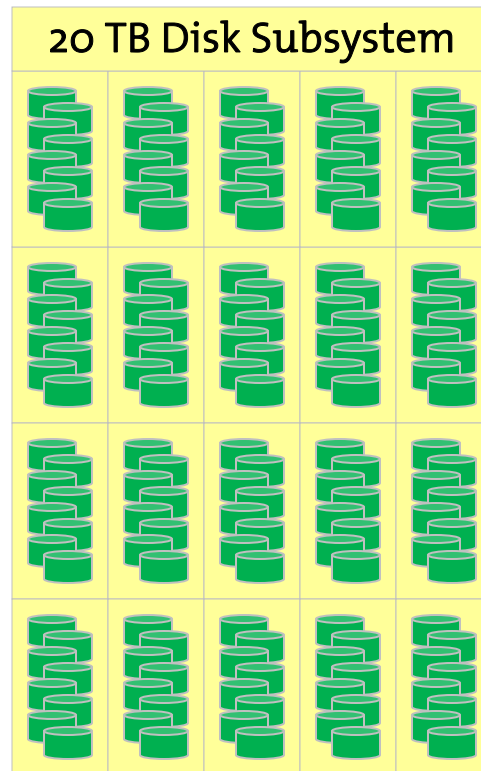
IntelliMagic

Access density

Number of I/Os per second per GB

Storage Intelligence

# Existing single-tier FC solution



Horizontal Storage Groups spread the load over all the array groups

10000 I/Os per second

Average access density = 0.5

Requires using 146 GB 15K RPM HDD

We need a total of 20 array groups

Total = 160 HDD

Note that for simplicity we assume that every front-end I/O generates one back-end I/O.



IntelliMagic

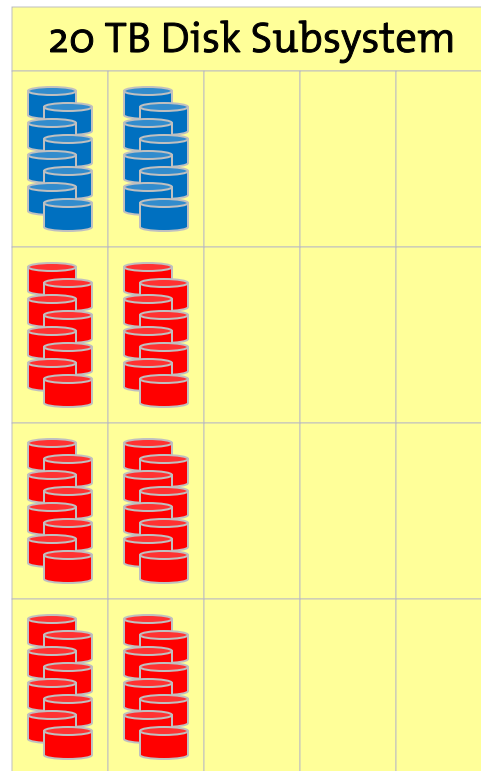
64/160 = 40%  
HDD required

For a large DSS  
a complete I/O  
frame may be  
saved!

Storage Intelligence



SSD / FC mix



10000 I/Os per second  
10% handling 90% of the load

2 TB = 9000 I/Os per sec Access  
Density = 4.5  
2 x 146 GB SSD array groups

18 TB = 1000 I/Os per sec  
Access Density = 0.06  
6 x 450 GB 10K RPM array groups

We need a total of 8 array groups

Total = 64 HDD

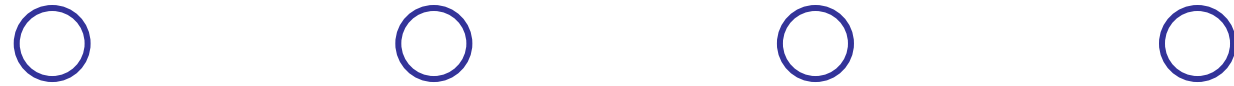


## What did we achieve?

- 64 drives rather than 160 drives
- Small number of SSDs handle most of the load  
25% of drives take 90% of load
- Can use large FC disks for what is left
- Much better performance: 90% of load at SSD speed



IntelliMagic



## How to justify SSD / EFD to yourself and to your manager



## Cost and Performance

- 
- 
- 
- 
- *Since SSD technology provides much better performance, any volume / data set is going to be a candidate when the price is right.*
- Right now, SSD is still expensive
- Consumer prices are dropping quickly, but ...
  - Intel 64 GB SLC costs 649 euro (Oct 7, 2009) a piece
  - Intel 160 GB MLC costs 379 euro a piece
  - 600 GB FC drive costs 618 euro in the same on-line shop
- So .. you still need to think about what data to move to SSD



## Reasons

### Business oriented:

- Your application needs to go faster (and disk access is part of the problem)
- Your batch window needs to be reduced (work files)

### Internal IT oriented:

- You use short-stroked disks for active databases
- You are tight on space in your data center
- You want to gain experience with the new technology



# How to pick your candidates?

What is the objective when implementing SSD

- Improve storage performance?
  - Find highest potential response time improvement
- Reduce cost: smaller and less expensive configuration?
  - Find volumes with highest back-end activity
  - May include very active sequential data sets

Note that in most cases you will get both benefits, a configuration with SSD will be both smaller and faster.



# ○ ○ ○ ○ Improve storage performance

Find highest potential response time improvement

- Data transfer on Fibre/FICON remains the roughly same
- Contribution from read-misses will disappear
- Contribution from synchronous copy will **not** change

So look for data sets / files with high read-miss activity

**The Good:** Maximum reduction in response time

**The Bad:** Write intensive workloads will continue to generate high HDD load



## Smaller configuration

Eliminate as many physical I/Os as possible from HDDs

- Find most active back-end activity on the spinning disks as much as possible
- Implement SSD and higher capacity HDDs for what is left

**The Good:** Exploits high I/O potential from SSD  
Smaller and faster configuration

**The Bad:** May not reduce (acquisition) cost just yet



# Migrate Data sets, devices or storage groups?

(presented in z/OS context)



## Data Set Level Selection

- 
- 
- 
- 
- Pick your best candidates based on SMF 42
  - IBM Apar OA25559 provides 'read miss disconnect time'
  - Or use physical I/Os based on cache statistics from SMF 42
- Results in very efficient use of SSD space
- Will need to review regularly or change SMS rules to keep right data on SSD
- Work data sets that cause a lot of activity harder to identify and manage
- Best suitable if you want to move 1% of your storage to SSD



## Volume Level Selection

- Good RMF data available to identify response or back-end activity candidates
- Analysis automatically covers temporary and permanent data sets
- Especially with back-end activity based approach, work volumes will be best candidates!
- Need to consider storage groups – active data sets may float around within storage groups



# Storage Group Selection

- Largest entity, will need more SSD ranks to have impact
- Workload profile for storage group will not change much, so 'safe' way to migrate
- No need to change SMS rules, simple implementation
- Applicability depends on storage group design and SSD cost
- Best when moving at least 10% of your data to SSD

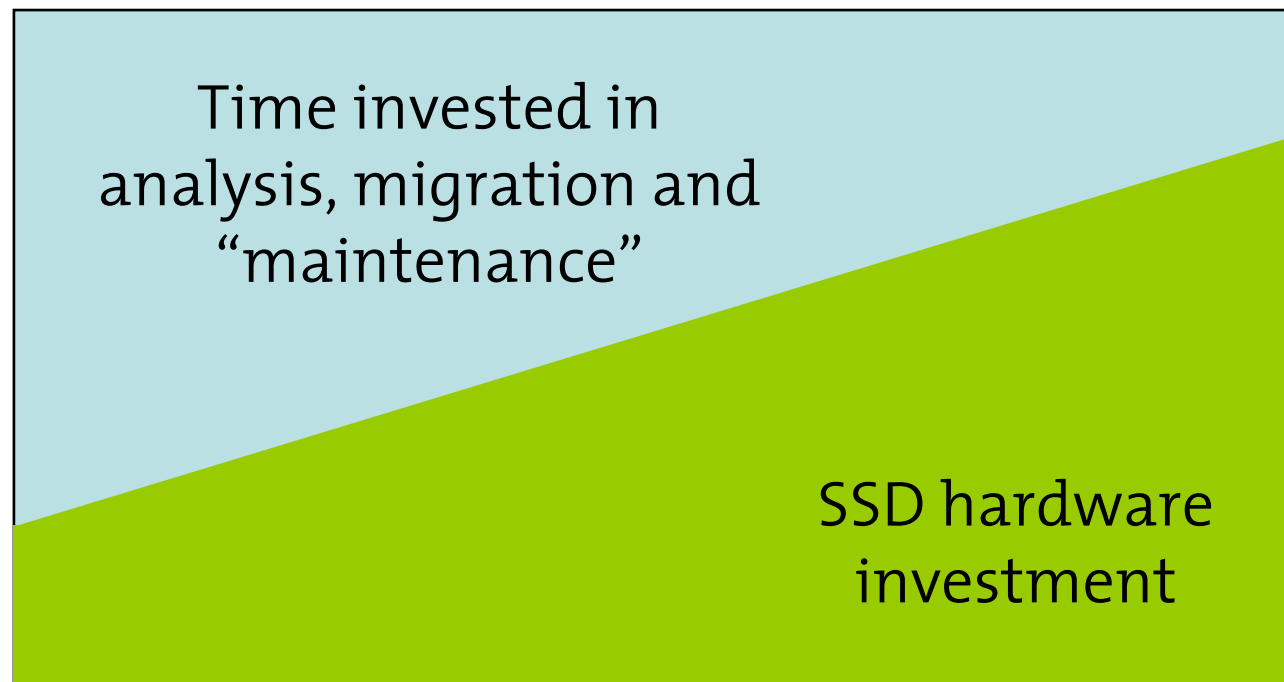


IntelliMagic

Storage Intelligence



## Hardware vs. people cost



Dataset/File

Device / LUN

Storage group



# Sample migration study



○ ○ ○ ○

# Large Financial Institution

z/OS installation

Looked at 1 Sysplex with

- 15,000 zSeries volumes
- 700 array groups, currently mostly (90%) 146GB/15 K RPM, some 73GB/15K RPM and 300 GB/10K RPM.
- 10,000,000 volume activity records processed (full week of RMF data for entire sysplex)

Study done was done with IntelliMagic Storage Performance Management Products (RMF Magic and Migration Advisor)



## Objective

- Perform volume level analysis to identify what portion of the very active data can be targeted to SSD with a small number of drives
- Studied adding 4, 6, 8, 16 or 32 SSD array groups, i.e. up to 5% of the storage.
- Goal is not to improve response time, but the reduce component count, i.e. use SSD with higher capacity HDDs.



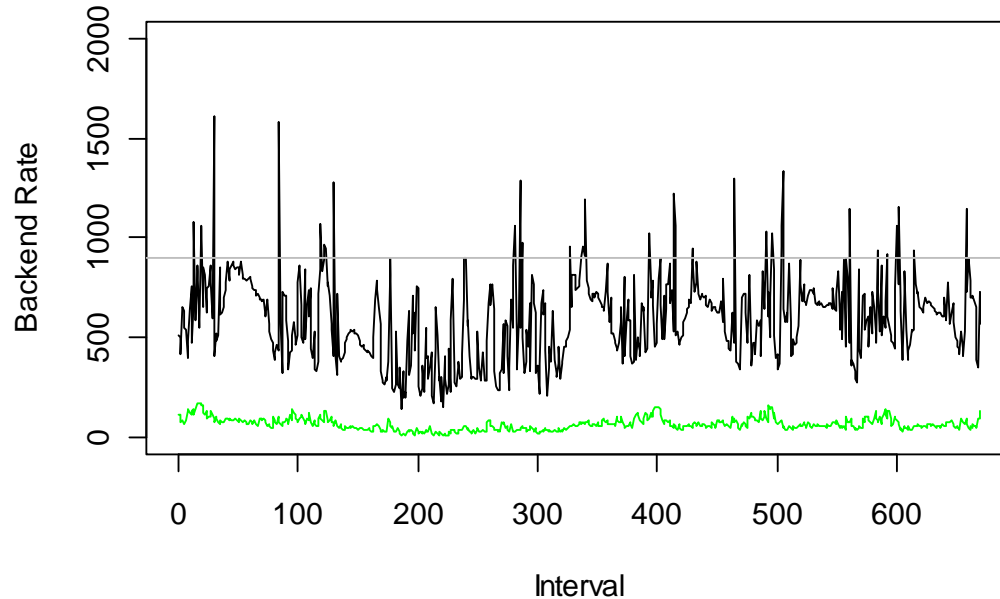
## Logic

- Find best logical volume -> array mapping such that 15K RPM array groups
  - Do not do more than 700 disk ops per second on average
  - No more than 5% of the intervals has an array group that exceeds 1000 disk ops per second
- Define 'growth potential' as growth that is possible while still satisfying the above criteria
  - When the average is 350, the growth potential is 100%
- Compute how 'growth potential' for HDDs changes as we add SSD ranks



Without SSD: all 146 GB / 15K RPM

Peak and Average



Black: Highest activity for any array group

Grey: 95% percentile

Green: Average for all array groups

Base line situation



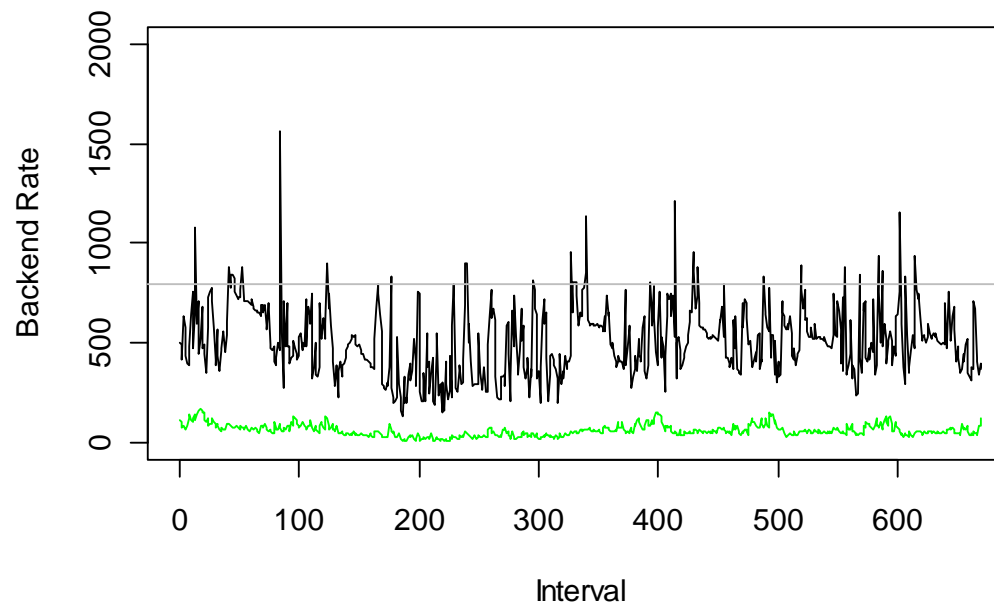
IntelliMagic

Storage Intelligence



## 4 SSD array groups added

Peak and Average



Black: Highest activity for any array group

Grey: 95% percentile

Green: Average for all array groups

4 SSD groups added, pick most active volumes from HDDs, chart shows HDDs



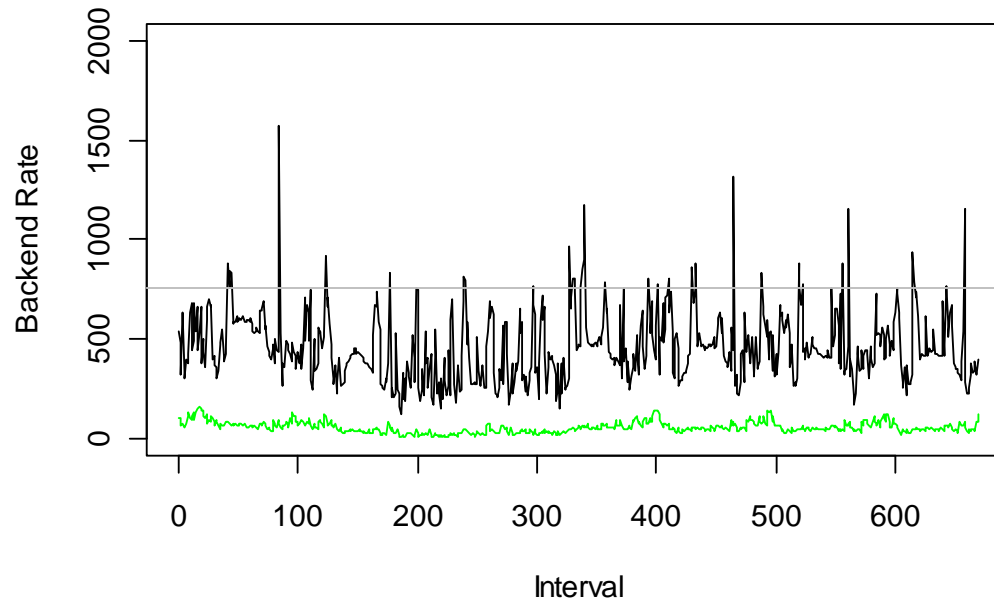
IntelliMagic

Storage Intelligence



## 8 SSD array groups added

Peak and Average



Black: Highest activity for any array group

Grey: 95% percentile

Green: Average for all array groups

4 SSD groups added, pick most active volumes from HDDs, chart shows HDDs



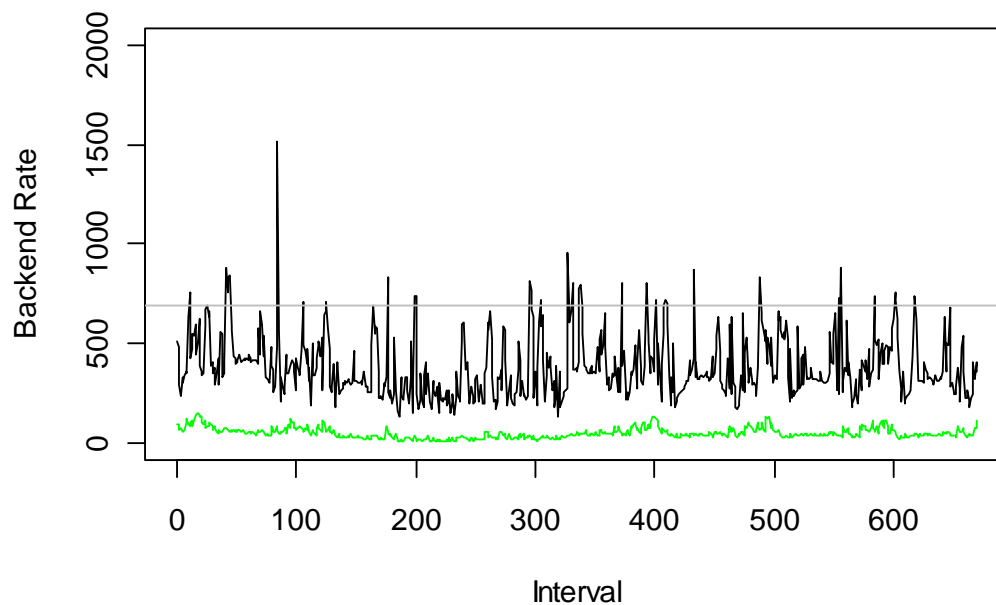
IntelliMagic

Storage Intelligence



## 16 SSD array groups added

Peak and Average



Black: Highest activity for any array group

Grey: 95% percentile

Green: Average for all array groups

4 SSD groups added, pick most active volumes from HDDs, chart shows HDDs



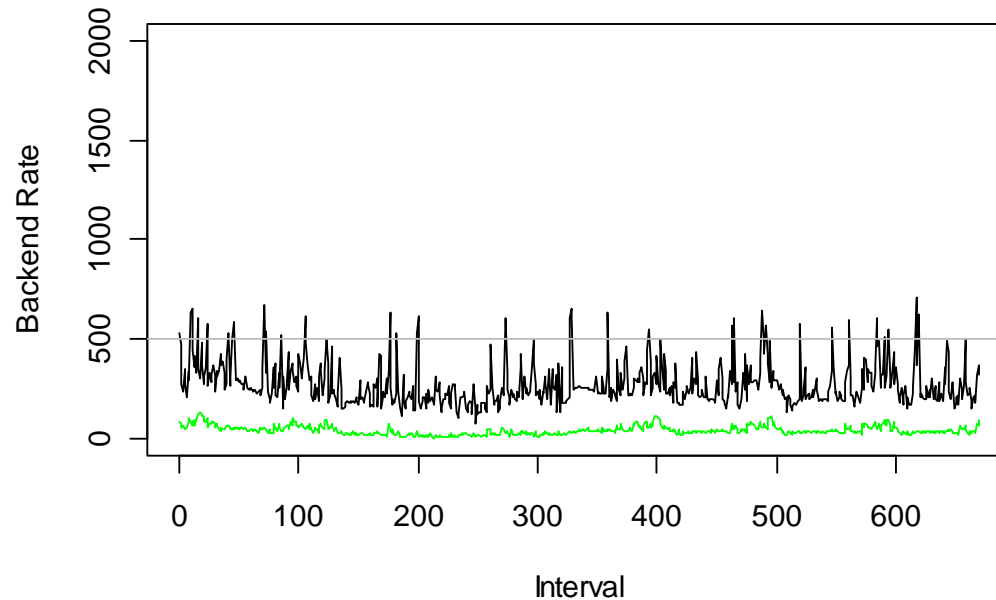
IntelliMagic

Storage Intelligence



## 32 SSD array groups added

Peak and Average



Black: Highest activity for any array group

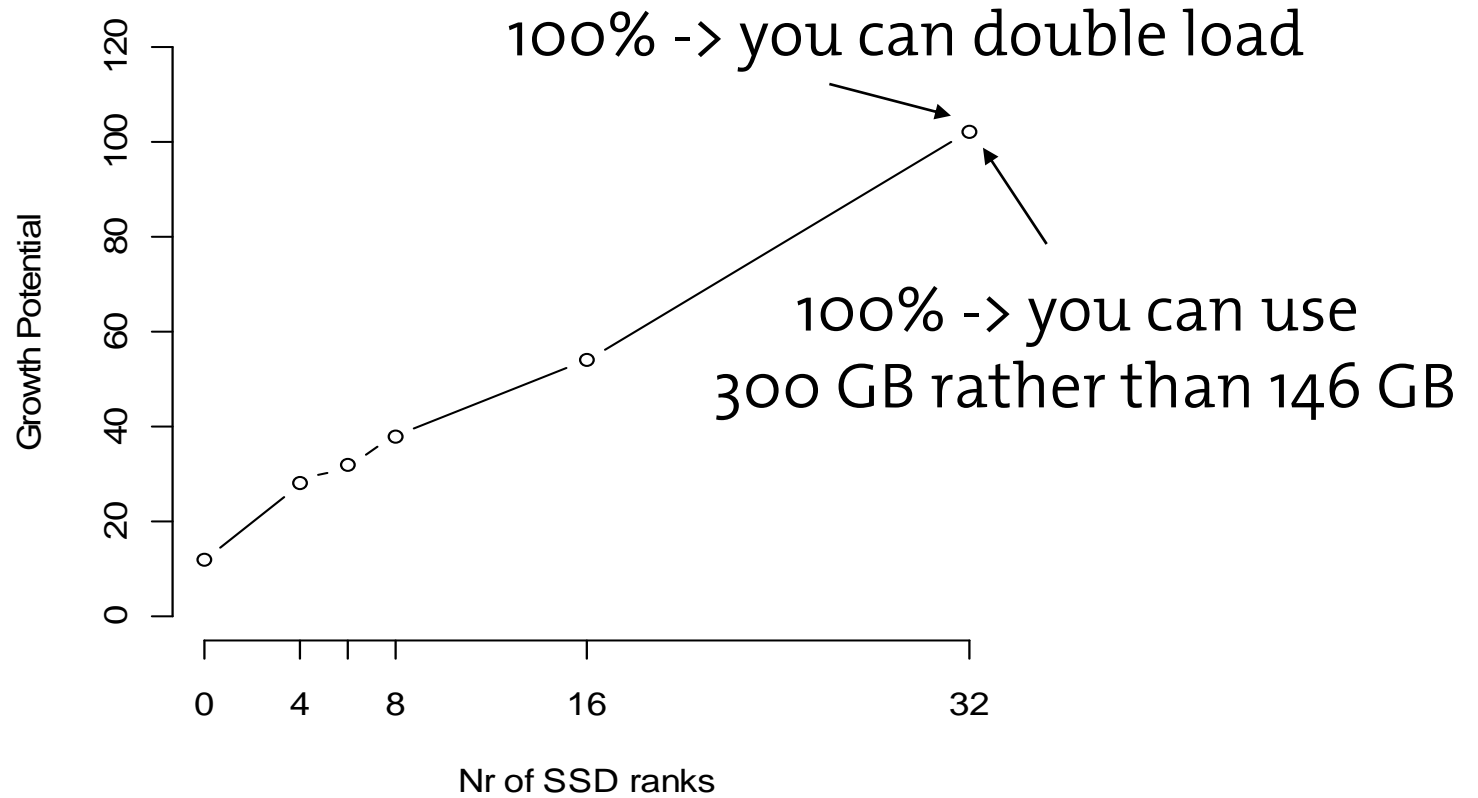
Grey: 95% percentile

Green: Average for all array groups

32 SSD groups added, pick most active volumes from HDDs, chart shows HDDs



# Growth Potential



How much extra work can the HDDs do?



## Summary

- By moving 5% of the data to SSD the remaining data can be place on 300 GB rather than 146 GB drives
- Reduction from 700 to 383 ranks
- 32 SSD ranks will handle about 42% of back-end workload!
  - this represents only 18% of the front-end I/O rate
- Resulting configuration will be faster
- May not be cheaper today, but will be cheaper very soon



IntelliMagic

Storage Intelligence



## Is this real?

- Our study is optimistic because
  - We look at one week and created an optimal configuration for that week. Another week may be different
  - We did not consider storage groups; data that moves within storage group may move off SSDs
- Our study is pessimistic because
  - We did not use any knowledge about workloads or data sets, cherry picking key data sets may help a lot
  - Our comparison is with an ideal starting point, actual configurations are less balanced, and hence more improvement is possible



Storage Intelligence

© IntelliMagic 2009



# Thank You

Questions?