



HRL Storage Power Management

Power Modeling of Storage Systems

Miriam Allalouf

Haifa Research Labs, Storage Power Management

Joint work with Ronen Kat, Dalit Naor, Kalman Meth, Yuriy Arbitman

5/19/08

© 2007 IBM
Corporation

Outline

- 'Storage Power Management'
 - Introductory Description and Trends
 - Problem space – list of items
- Power Estimation and Power Modeling
 - Why do we need power estimation instead of measurement?
 - What is storage power modeling?
 - Capacity-based power modeling
 - Utilization-based power modeling
 - Workload-dependant power modeling
 - Method to estimate power
 - High-end controller modeling → RAID array importance
 - Disk Technology and Modeling
 - Calculating disk activities
 - Usage : offline tools and capacity planning , On-line systems
 - Results
 - Summary and Future Work

Data Center Energy Problem

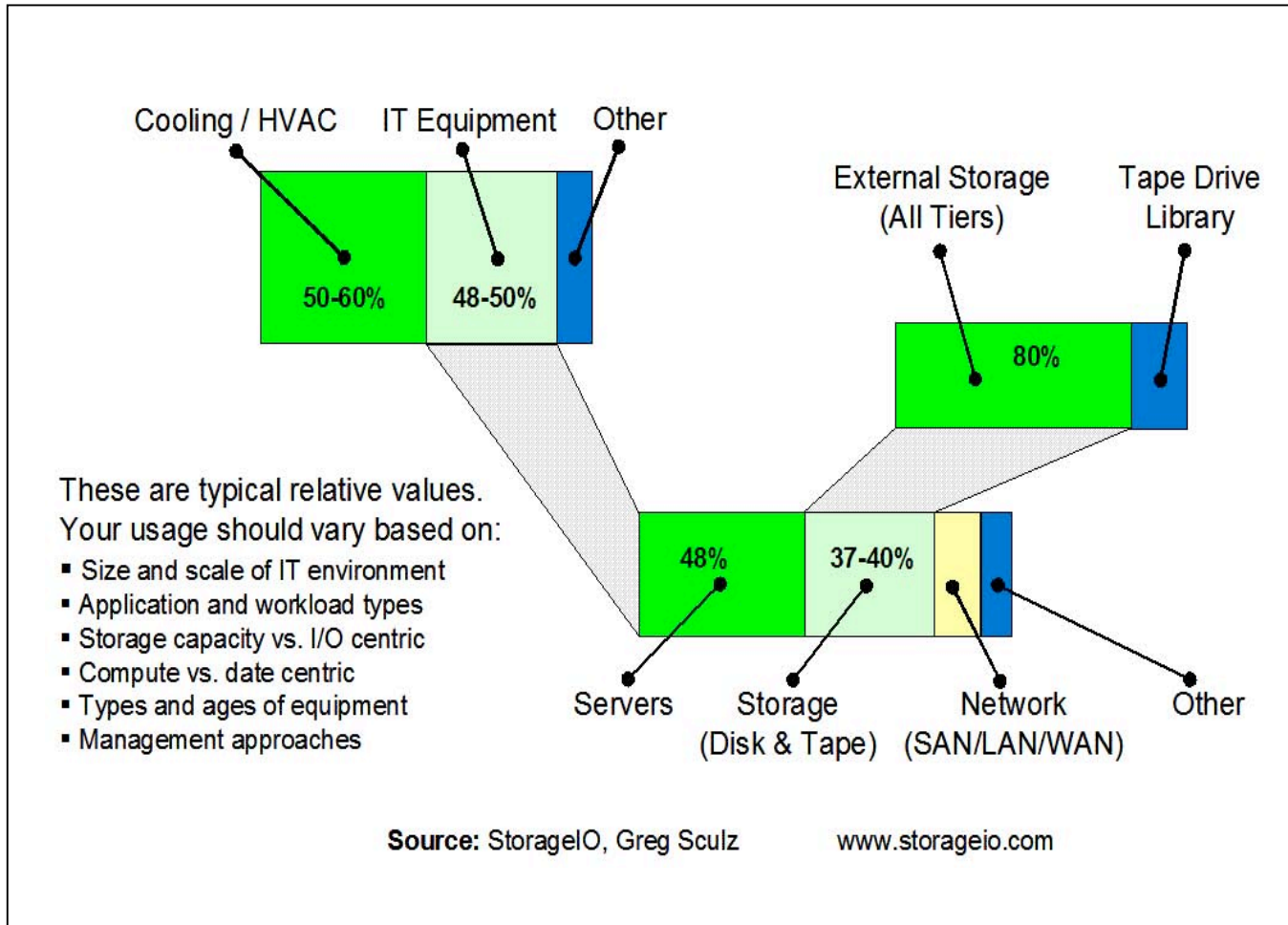
General Trend for all products:

- Rising Heat load per product: Watts per Equipment Square feet
 - Communication (extreme density), compute Servers, storage servers, standalone workstations and even tape storage (low)
 - Rising Energy costs
 - Higher Power consumption
 - High bills
-

The expanding data center where lot of power consuming products are concentrated together has to take care of:

- Increasing Equipment Density
cooling → hitting → cooling
→→
Non – reliable system
- Electricity costs that becomes part of IT budget

Data Center Energy Distribution



World-Wide Trends: Expanding Digital Storage

- **New Application types;**
- **categorized to content depots, Fixed content, replicated data, traditional business data**
 - Storage increase
 - New Storing Medias
- **Increasing Server/Storage Performance Gap**

Storage Power Management – Problem Space

Power Management is not “Power Saving” only ... and “Power Saving” can have many aspects

- Planning tool for the design of power-aware system
- **Suggested** Power –aware Technologies
 - Heterogeneous power consumption media:
 - disk (FC, SATA), tapes
 - Use SSD or Flash memory instead of disk
 - Use mobile disks in enterprise environment.
 - Too Short idle periods. Less reliability.
 - Smaller size. Limited performance
 - Dynamic RPM: Uses the fact that the spindle power is quadratic to its speed.
 - Not available at the moment.

Storage Power Management – Problem Space (cont 1)

- Massive Array of Idle Disks (MAID)
 - Caching-based MAID (Colarelli and Grunwald)
 - Migration-based MAID (Pinheiro and Bianchini, Zhu et al.)
- Power-aware HW capabilities
- Power reporting capabilities:
 - Disk, Enclosure, Logical volume
- Power-aware data management and optimization algorithms
 - Disk Spin-down to conserve power ().
 - idle periods: many, but too short.
 - Break-even threshold time.
 - Power-aware data migration: maximize in order to shutdown
 1. the number of idle disks
 2. The time between two consecutive accesses to the same disk.

Storage Power Management – Problem Space

Power Management identified with “Power Saving”

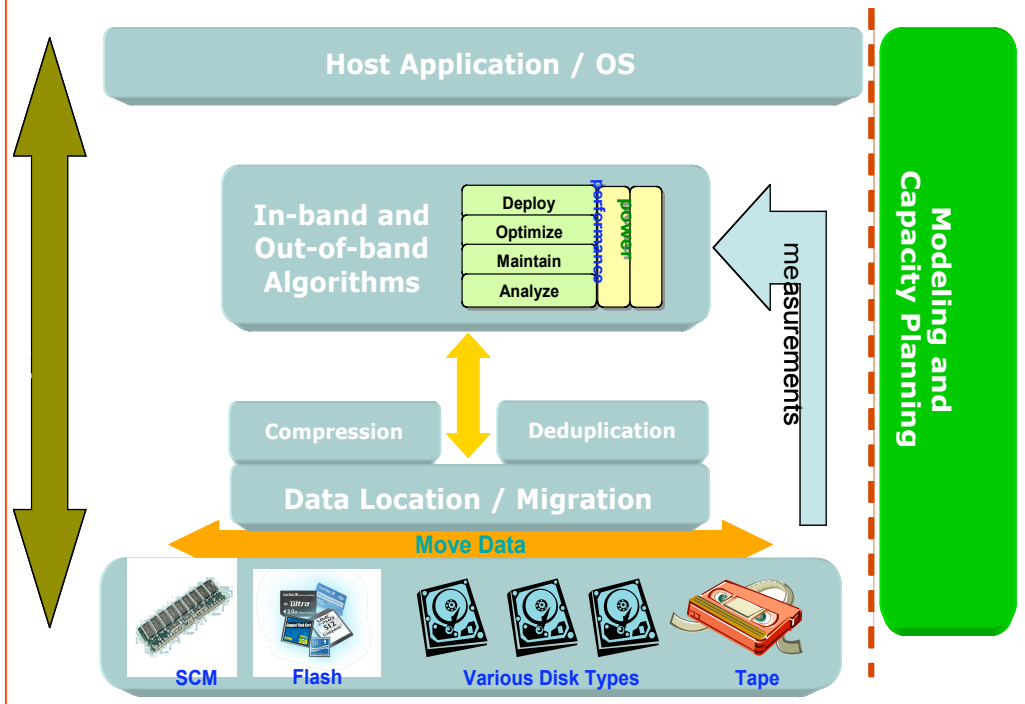
- Caching Objective:
 - increase the idle time between disk requests.
 - Methods:
 - Write caching,
 - always prefer evicting blocks from active disks.
 - another is allocating more cache space for inactive disks than active disks.
 - Pre-fetching
 - Cost-aware caching.
- Reduce Redundancy for Power saving.
 - Shutting down redundant disks. (RAID5, RAID6, erasure codes)
 - Use NVRAM for intermediate storage.
 - Change the number of RAID strips:
 - Bigger strip size: more performance.
 - Smaller strip size: less power.

Power-Aware Storage Road Map

To appreciate benefits and optimize appropriately - required:
 Accurate Power Estimation
 Power terminology



- Modeling
- Measurements and benchmarks
- Metrics and standards



Outline

- 'Storage Power Management'
 - Introductory Description and Trends
 - Problem space – list of items
- Power Estimation and Power Modeling
 - Why do we need power estimation instead of measurement?
 - What is storage power modeling?
 - Capacity-based power modeling
 - Utilization-based power modeling
 - Workload-dependant power modeling
 - High-end controller modeling → RAID array importance
 - Disk Technology and Modeling
 - Calculating disk activities
 - Method to estimate power Usage : offline tools and capacity planning , On-line systems
 - Results
 - Summary and Future Work

Why power estimation instead of measurement?

- Capacity planning and prediction tools must use external power data and extrapolate
- On-line system can not attach meter per each disk or even enclosure.
- Logical entities can not be measured physically

Power Modeling Goals:

Gain knowledge of the Actual Consumed Power

- Know the amount of the consumed Watts per each component in the system:
 - During a period of time
 - Given different stages of the activity.
- Example:
- Labeled 'Watt per GB' disk power consumption vs.
- Actual Watt per I/O given a workload characterization, idle periods and HDD utilization.

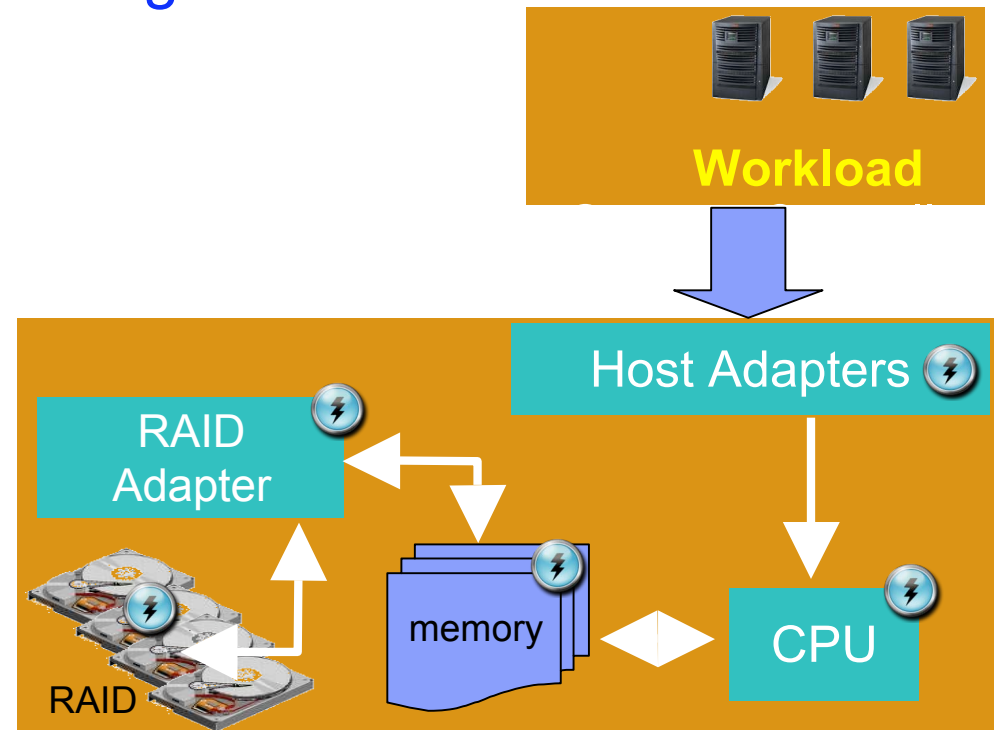
Modeling for Capacity Planning

Capacity Planning Tool

predicts the utilization of the various components of storage systems given certain configurations.

Power Modeling for Capacity Planning

- Focus on high-end controller (not simple RAID)
- Model the power consumption of the different storage controller components for various workloads characteristics.



How different Performance Model from Power Model?

- **Performance metrics:** Response Time, Throughput of I/O per second
 - System is modeled only when I/O occurs
- **Performance Modeling** use I/O characteristics at the host level and translate
 - Parameters that affect time: IO rate, Read/Write Ratio, sequential/Random ratio, Cache hits.
- **Power Modeling** has to model all system states: 'idle' and 'working'
- **Power Modeling** has to reflect the power consumption upon various working levels, power states and configurations

Power Metrics

- **Power Metrics can be:**

- Watts per GB ?

Capacity Power modeling

- Or Watts per system utilization – at various levels of I/O rate ?

Utilization-level power modeling




- Or Watts per workload characteristics and IO rate?

Workload-dependant power modeling

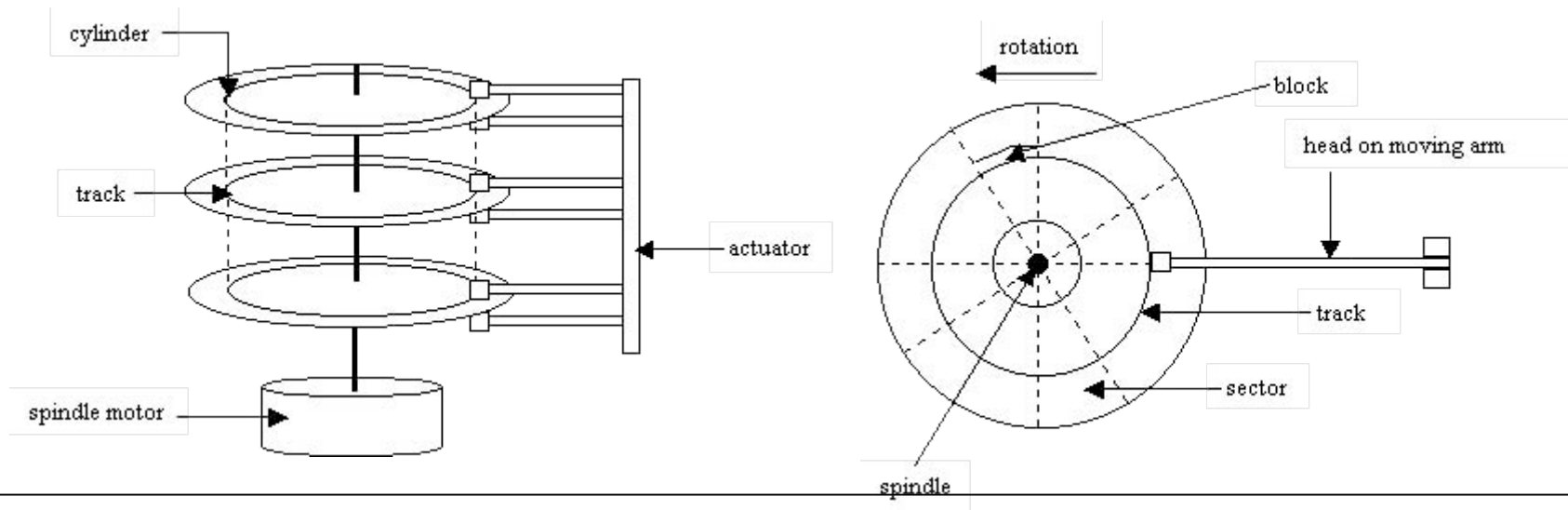
High-end controller power breakdown

High-end controller w/ 384 HDDs:

More than 58% of power is dedicated to disks drives

-  Identify the controller power consuming components
-  Focus on the RAID arrays and disks
-  Next steps – other components: SMP, I/O, etc

Disk Technology



- Disk Characterized
 - RPM, I/O Interface, Capacity, Size, Seek time
 - Access Time (latency + seek),
 - Each feature implies other power consumption
-
- About $\frac{2}{3}$ of the power is for spinning the platters. NO power-aware intermediate modes
 - The platters are spinning even when “idle”.
 - About $\frac{1}{3}$ of the power is for serving I/O requests:
 - the voice-coil actuator seek operations and the
 - read-write magnetic head data transfer

Seagate - Cheetah 15K.5 FC Taken from Seagate product manual

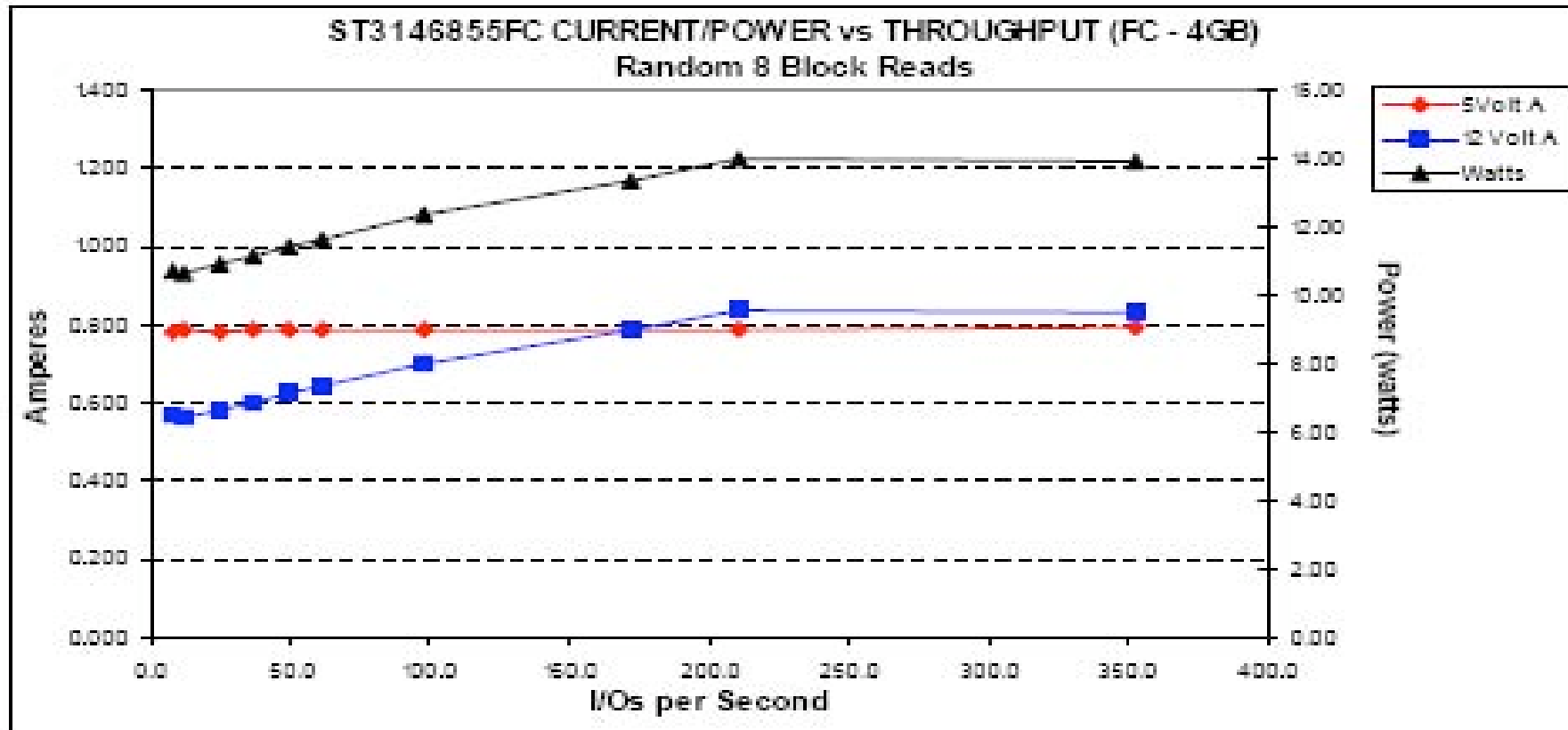


Figure 8. ST3146855FC DC current and power vs. input/output operations per second at 4 Gbit

$$\text{Watts} = 12V * I (\text{Amp}) + 5V * I (\text{Amp})$$

Disk Power Consumption and Performance counters

$$Power(disk_i) = Static_{power} + Dynamic_{power}$$

12V

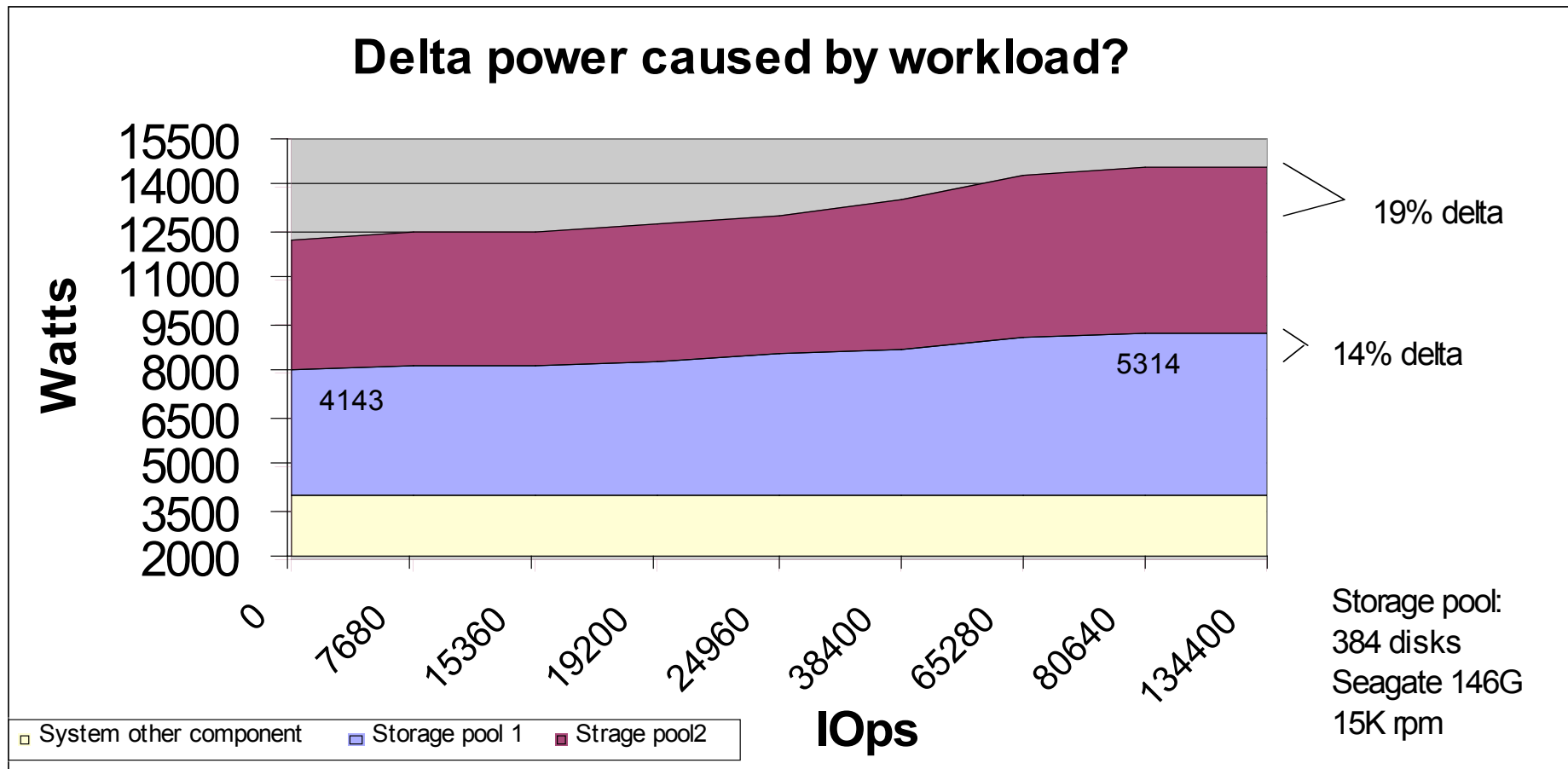
5V

$$Static_{power} = Power(spindle) + Power(electronics)$$

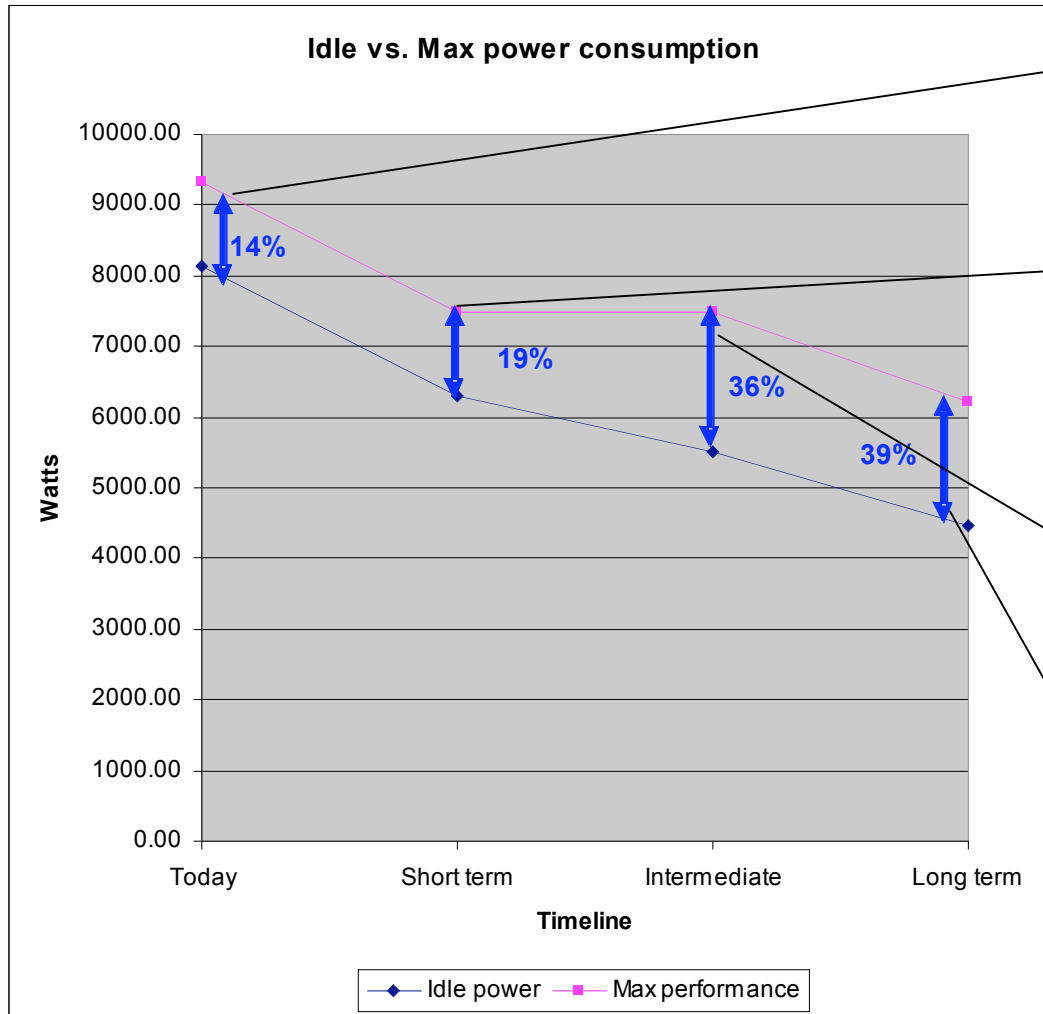
$$Dynamic_{power} = \sum_{I/O} [Power(electricalActivity) + Power(mechanicalActivity)]$$

Disk Power Consumption can be known
 BUT it is hidden underneath Controller's virtualization layers
 There are NO performance counters that reflect those activities.

Why should we model utilization if power delta is so small?



Near future: utilization level will affect storage power



384 146G 15k drives
Random Read Workload

First 'power saving' steps:
20% less in 'system idle and max'
25% less in 'disk idle'

Leveraging server technology:
40% less in 'system idle'
20% less in 'server max'
25% less in 'disk idle'

Leveraging storage technology:
SSD and power management tools
40% less in 'system idle'
50% less in 'disk idle'
20% less in 'disk max'

Workload-dependant power modeling

<u>EXAMPLE</u>	Scenario 1	Scenario 2
I/O rate	1500 I/Ops	1000 I/Ops
Read %	70%	30%
Read sequential (out of reads)	0%	0%
Read hit %	50%	50%
Write %	30	70%
write sequential (out of reads)	50%	50%
Write efficiency	0	0
Seek % (or seek per I/O?)	33%	33%
HDD utilization	50%	56%
Response time	4.3	2.1
Power	?? = ??	

Seagate - Cheetah 15K.5 FC Taken from Seagate product manual

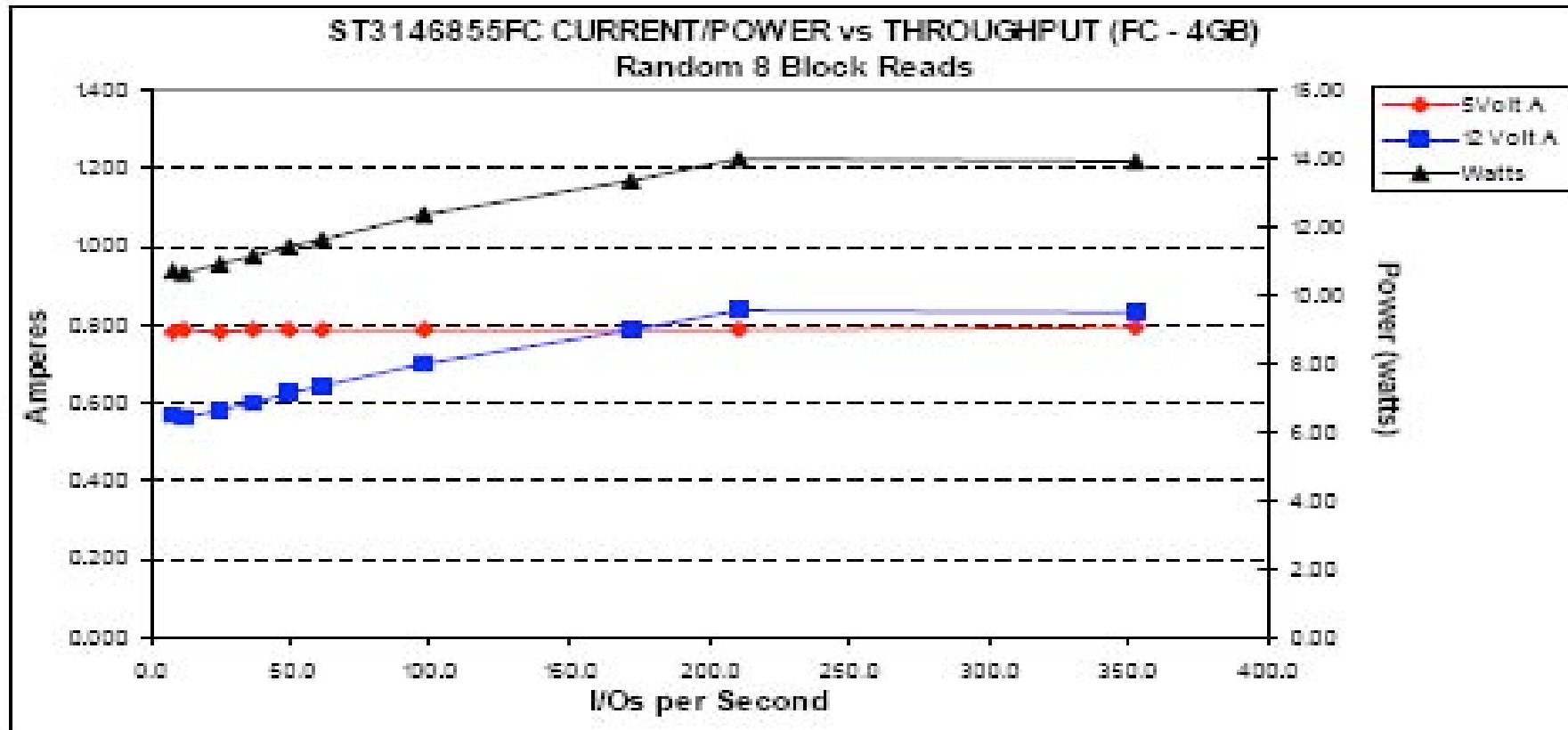


Figure 8. ST3146855FC DC current and power vs. input/output operations per second at 4 Gbit

$$\text{Watts} = 12V * I (\text{Amp}) + 5V * I (\text{Amp})$$

Method for Storage Power Modeling

- Assumes host workload characterization (as for performance modeling)
- Calculating Disk activity from an I/O workload
- Compose offline power consumption measurements
- Calculates Power in Watts upon idle+activity
 - A computation that estimates Workload-dependant power consumption using offline measurements.
 - This method can be inserted into capacity planning tool for power and performance prediction
 - This method can be used to estimate power in an on-line system

Validation Benchmarks

Test	Measured Values			Estimated Value			Percentage
	+12V(max)	+5V(max)	Power(max)	+12V(max)	+5V(max)	Power(max)	
Idle	6.285	5.7	106.452				
Write Benchamrk	8.275	5.955	132.153	8.76589	6.18	139.277	5.18%
OLTP benchmark	8.755	5.895	137.751	9.3787	6.18	146.8767	6.7%
Read Long transfer size benchmark	8.84	6.22	140.5	9.56	6.18	149.106	6.2%
Sequential read benchmark	5.935	6.915	108.3	6.285	6.18	108.8235	0.5%
Sequential write benchmark	6.005	6.81	108.6	6.285	6.18	108.8235	0.5%

The yellow columns hold the benchmark results

The orange columns hold the interpolated results

deviation

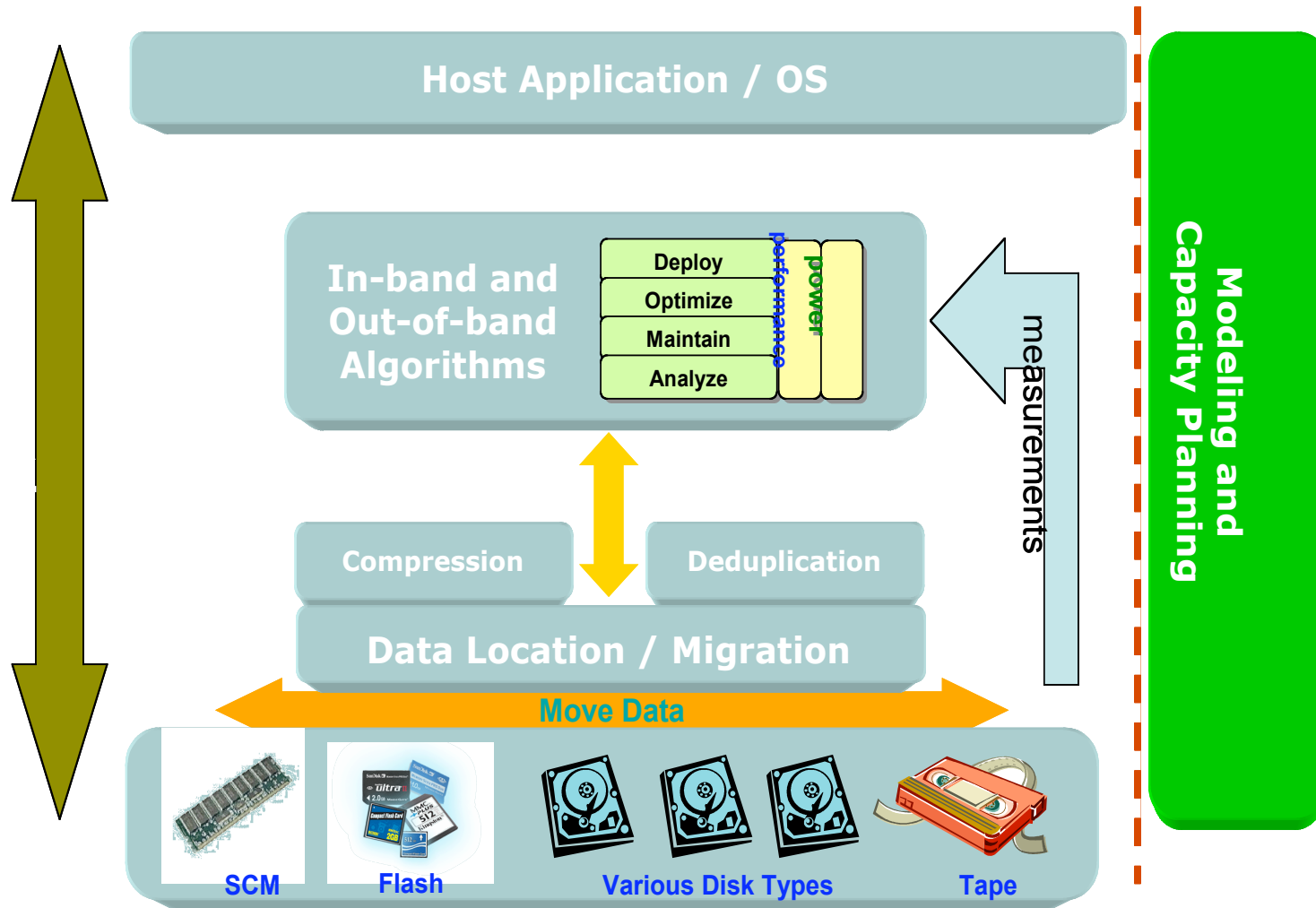
Related Works

- “Modeling Hard-Disk Power Consumption” Zedlewski, J., Sobti, S., Garg, N., Zheng, F., Krishnamurthy, A., and Wang, R. FaST 2003
 - Disk Energy Modeling and Performance Simulation Environment: get detailed disk I/O trace and interprets both the performance (using DiskSim) and the related power consumption of each I/O operation. An event-driven tool → compute the exact mechanical and electrical operation of the disk drive. The implemented method in Dempsey is impractical in enterprise systems for online systems and Impossible for offline system – due to lack of (future) trace data.
- Energy Management for Hypervisor-Based Virtual Machines, Jan Stoess, Christian Lang and Frank Bellosa. In Proceedings of the 2007 USENIX Technical Conference 007.
 - A method to estimate disk storage power by computing the time that each I/O request takes. The model considers the idle and active disk modes and their known associated power consumptions. Our approach estimates the power consumption based on the amount of disk operations and not on the time the operation takes. Note that disk operations may take the same amount of time, but consume different amount of power.

Power Estimation for the CPU

- Event-Driven Energy Accounting for Dynamic Thermal. Management. Frank Bellosa Simon Kellner Martin Waitz Andreas Weissel
- Run-Time Power Estimation in High Performance Microprocessors - Russ Joseph and Margaret Martonosi

Future Work



Thank You!