

Addressing the Stranded Power Problem in Datacenters using Storage Workload Characterization

January 30th, 2010

Sriram Sankar and **Kushagra Vaid**



GLOBAL DELIVERY



ONLINE SECURITY



INFRASTRUCTURE
SERVICES

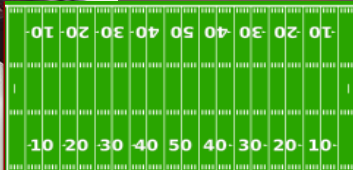


ENVIRONMENTAL
FOCUS

Microsoft Online Services

Across the company, all over the world, around the clock



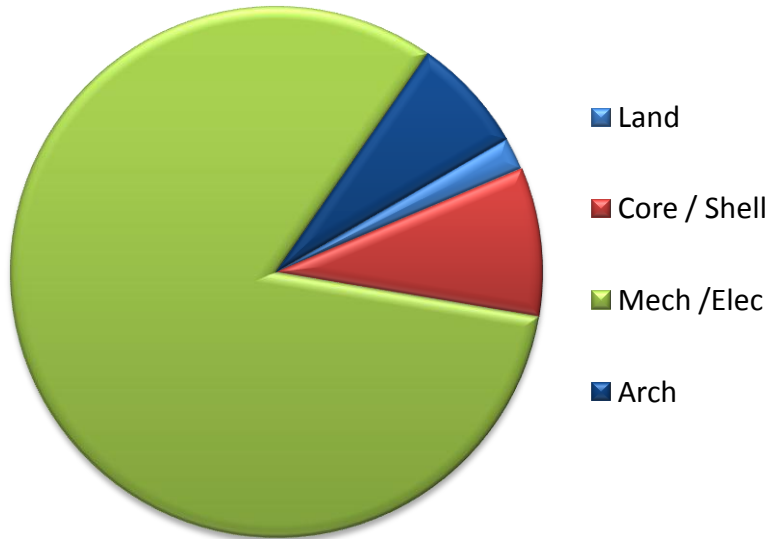


Large Datacenters can be
~**10 times** the size of a football field
and consume **10's of MW** of power



Cost of power in Datacenters

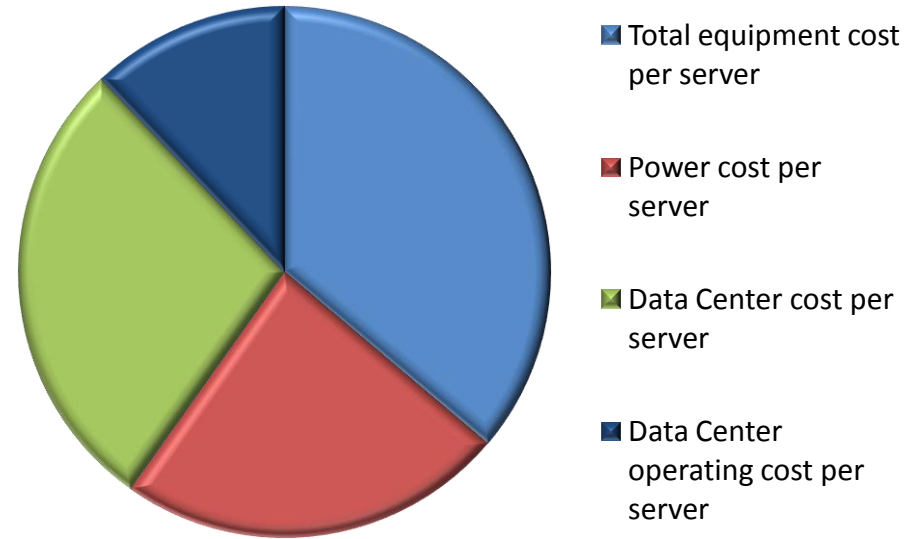
Datacenter investment breakdown



Datacenter capital cost: \$200-\$300M
\$10-\$20 for every Watt Provisioned

Where the costs are:
> 80% scale with power
< 10% scale with space

Basic 1U Server - 5 year TCO



Amortized Datacenter Infrastructure costs account for ~25% of the Server TCO

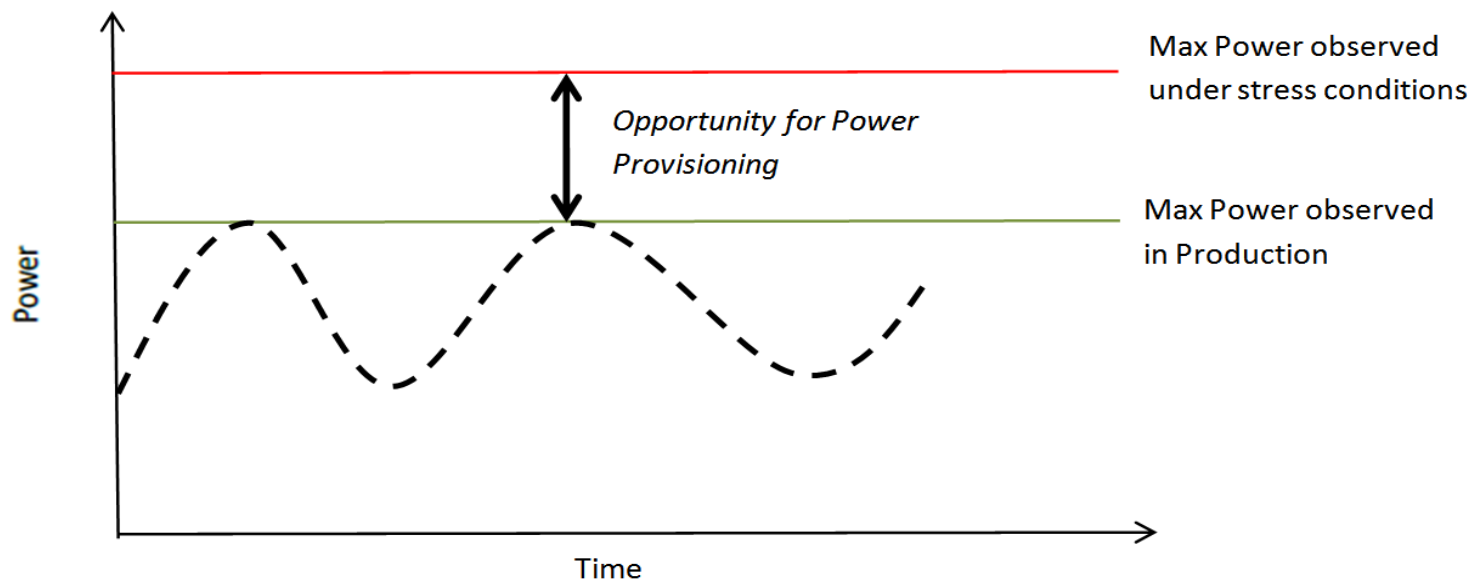
Server Power Consumption accounts for ~25% of Server TCO

Stranded Power Problem Statement










✦ Stranded Power = Allocated Power – Consumed Power

✦ What server power value should be used for datacenter power capacity planning?

- MaxPower under stress load? *Too high, power capacity is unutilized*
- MaxPower based on historical averages? *Risk of overloading datacenter circuit during peak activity periods*



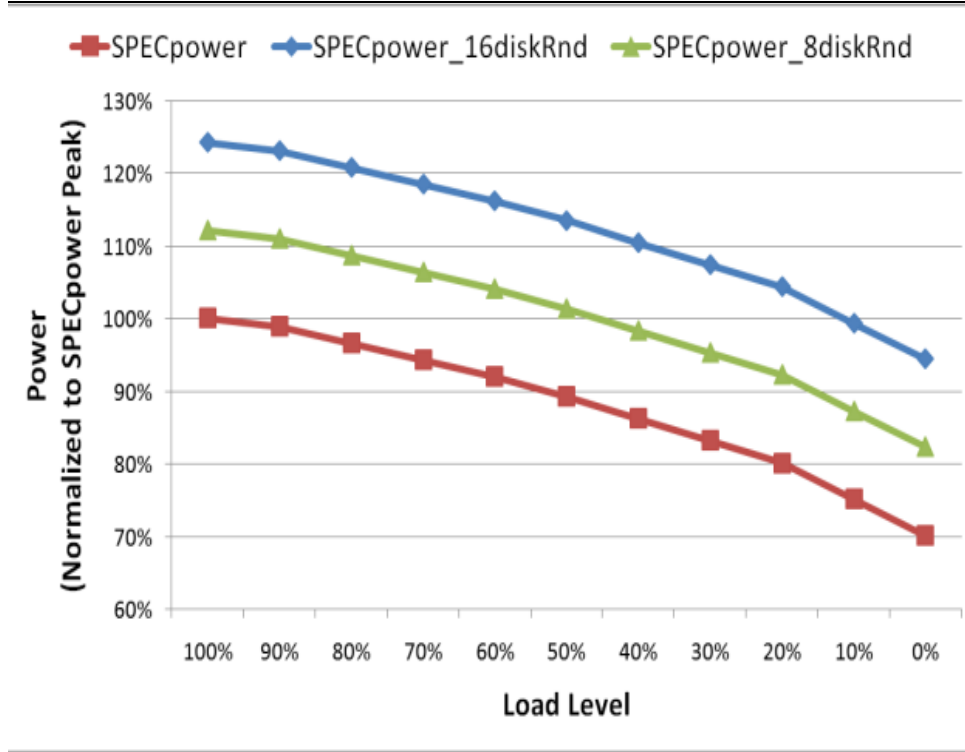
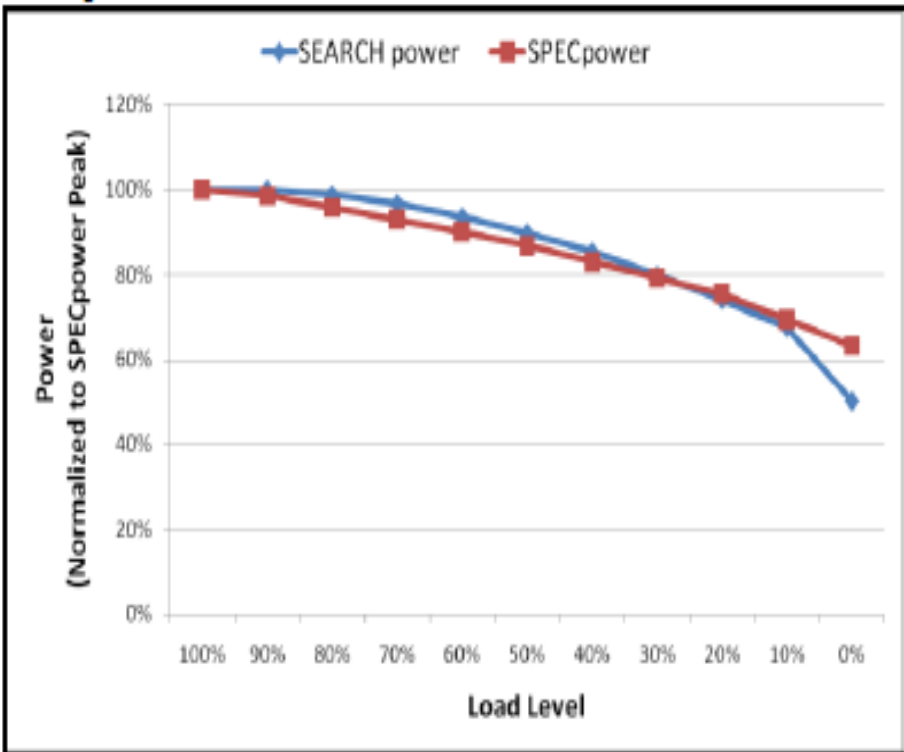
Perfect Power Provisioning is Challenging

-   Variety of workloads and server types
-   Production scenarios not reproducible in labs
-   Peak Vs Average power consumptions
 -  Large deviation between peak and average power
-   Datacenter Overload impacts Service availability








*SPECpower*_ssj2008 for server power estimation








- SEARCH power is fairly representative of compute-intensive workloads, however disk subsystem not represented
- Need a workload-driven methodology for accurately determining storage subsystem power










Focus area for this paper

-   Use SPECpower to determine power consumed by CPU+Memory subsystem
- 
- 
-   ***Devise methodology for determining storage power consumption for production workloads***
-  Use SPECpower + I/O stress tools to determine server power for datacenter capacity allocation

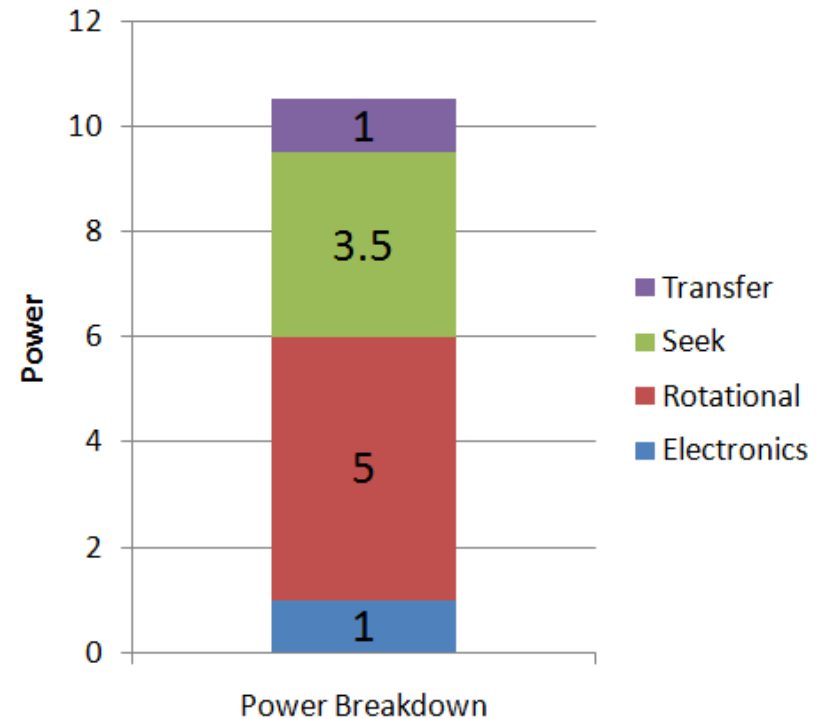
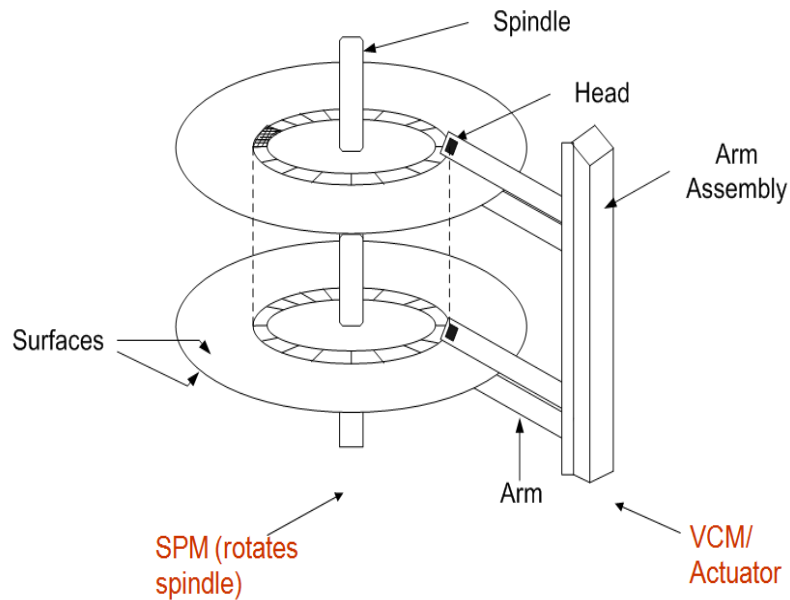
Our Proposed Methodology

-   Disk power characterization by exercising synthetic I/O patterns
-   Analyze production workloads I/O-request traces and generate corresponding I/O profiles
-  Estimate power provisioning value for datacenter capacity planning using SPECpower + I/O profiles

Our Proposed Methodology

-   Disk power characterization by exercising synthetic I/O patterns
- 
- 
-   Analyze production workloads I/O-request traces and generate corresponding I/O profiles
-  Estimate power provisioning value for datacenter capacity planning using SPECpower + I/O profiles

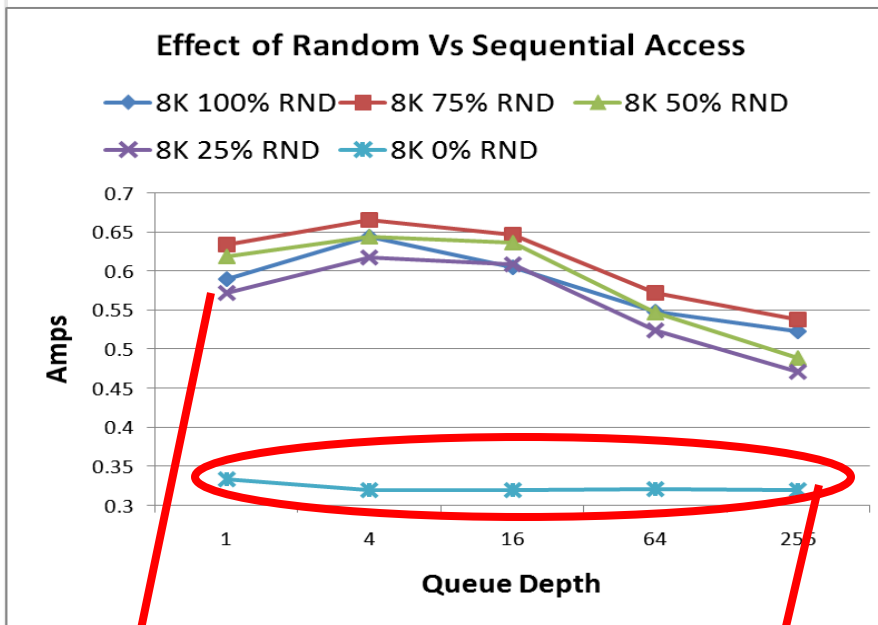
Disk Drive Power Primer



Disk drive power is determined by ...

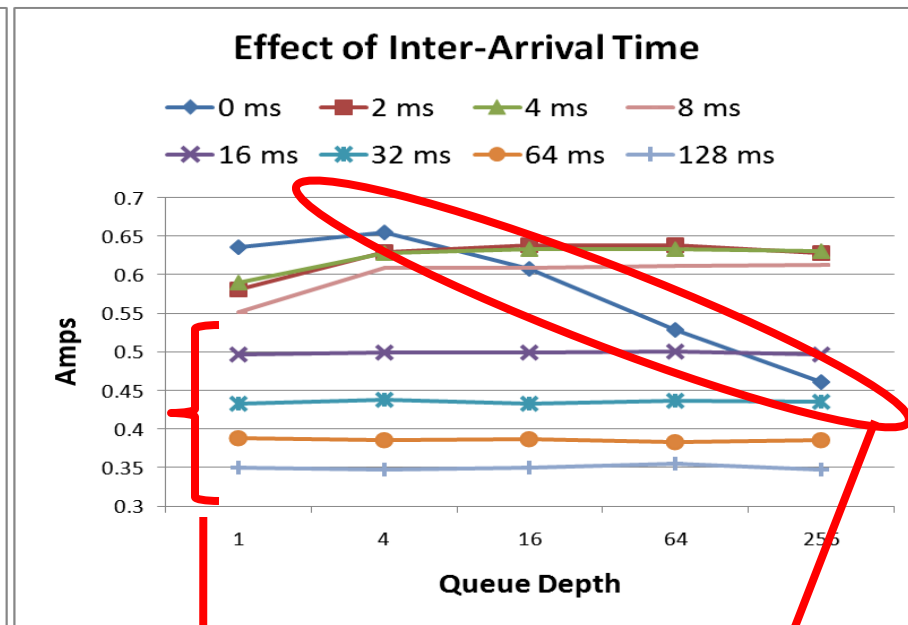
- Rotational motion from Spindle Power Motor (SPM)
- Seek activity driven by Voice Coil Motor (VCM)
- Electronics (disk controller)
- Data Transfer

Disk Power Characterization



Not much difference in power for variations in degree of I/O request randomness

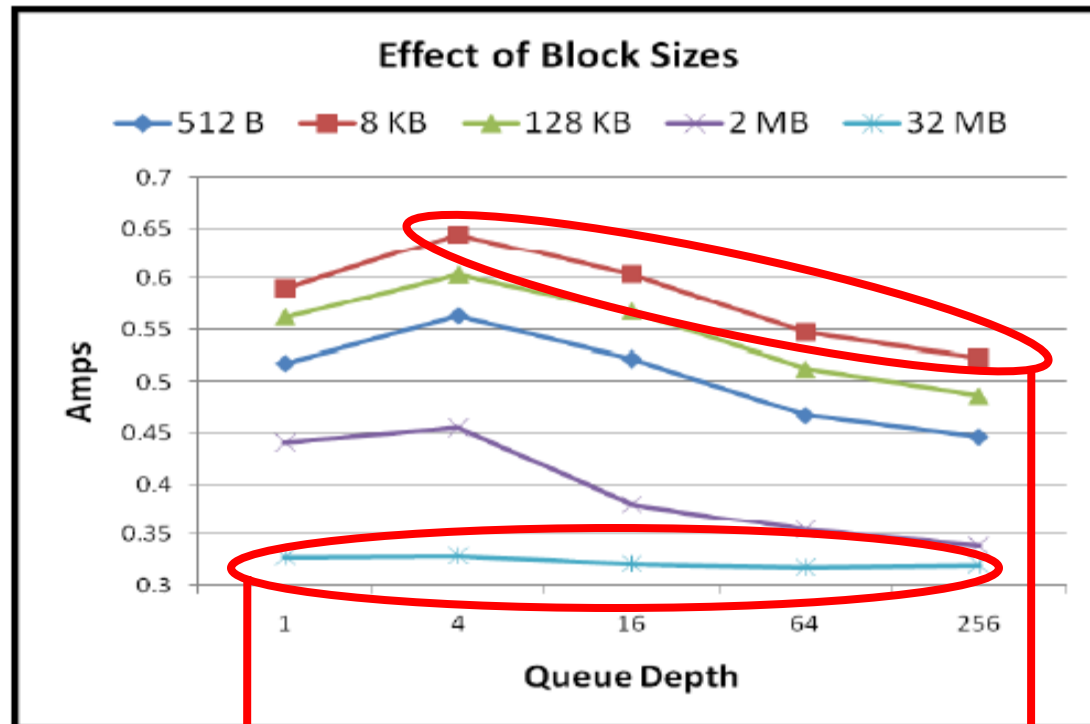
Sequential Access → No Seek Activity
→ Close to Idle Power



More I/Os at high arrival rates
→ Seek Optimizations by disk controller
→ Close to Idle power

Larger Inter-Arrival times → Minimal Seek Activity
→ Close to Idle Power

Disk Power Characterization



Larger block sizes → No Seek Activity
→ Close to Idle Power

8KB block size consumes the most power –
recall disk power is a function of transfer
size and seek activity

Our Proposed Methodology



- Disk power characterization by exercising synthetic I/O patterns

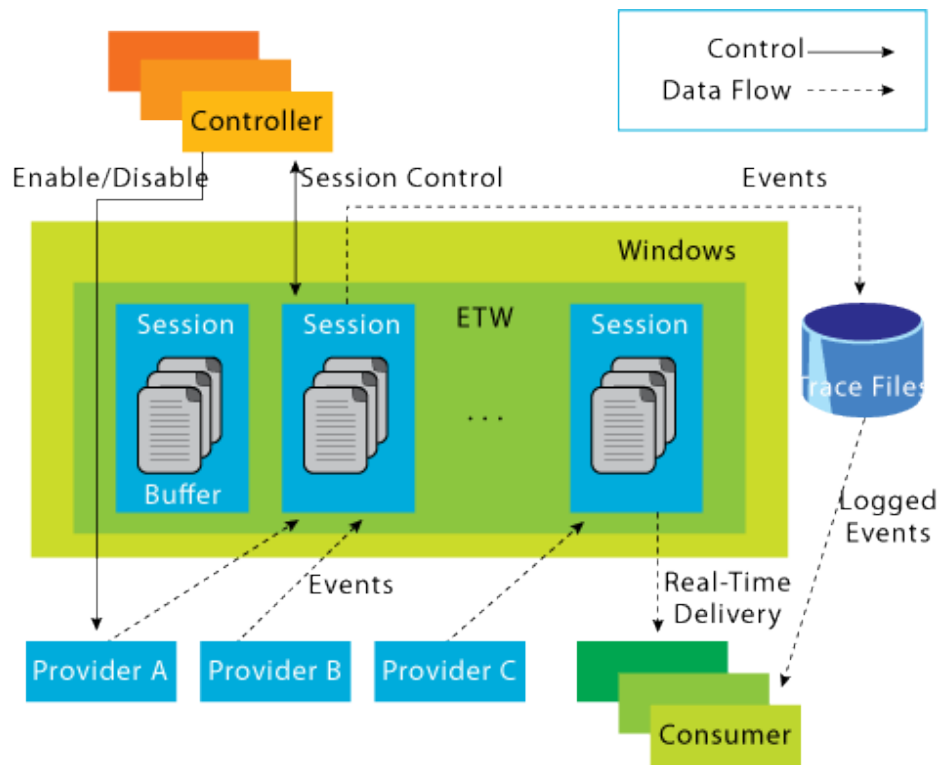


- Analyze production workloads I/O-request traces and generate corresponding I/O profiles

- Estimate power provisioning value for datacenter capacity planning using SPECpower + I/O profiles

Tracing Infrastructure - ETW

Event Tracing for Windows (ETW)



Examples of Events Captured

- Disk RD/WR Start, Completion)
- Timestamp of request
- Process/Thread id
- Request Offset
- Size of request in bytes
- Disk number as viewed by the OS
- Disk service time

Applications analyzed



Real Production Workloads

MAPS (Bing Maps)

– Texture and imagery repository



BLOB-DB (Windows Live Storage)

– Storage Tier hosting user content



Windows Live Spaces



Photos

SkyDrive: 25 GB of free online storage



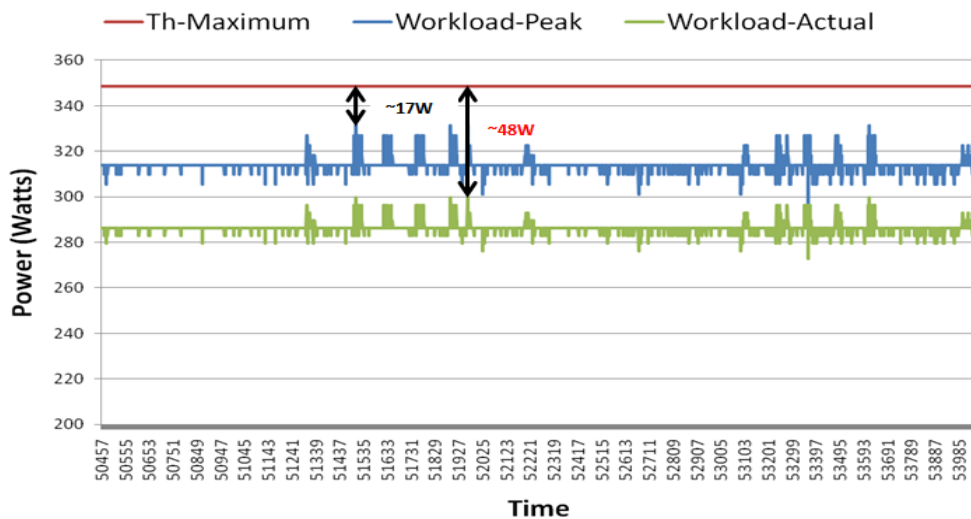
Deriving Power consumption using Trace-Driven Approach

Workload I/O profiles

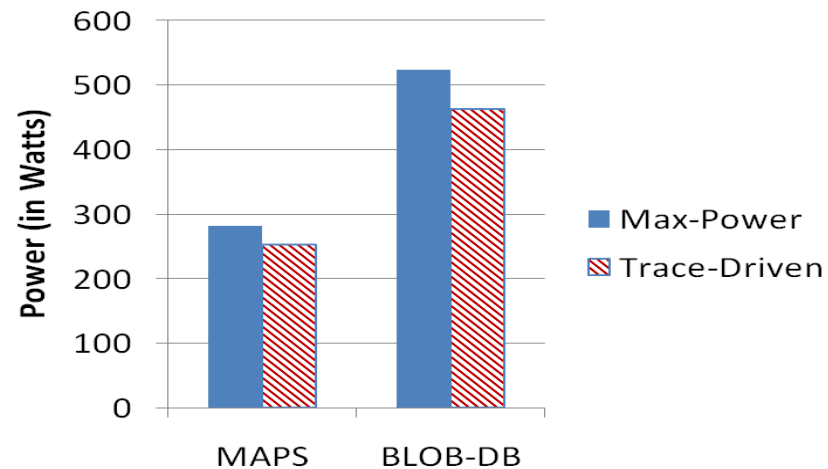
Workload	Randomness			RD:WR Ratio	Read					Write				
	Total Rd + Wr	RD	WR		4K	8K	16K	32K	64K	4K	8K	16K	32K	64K
BLOB - DB	89%	93%	77%	3.2		58%			2%		19%			
MAPS	27%	24%	96%	19.8	14%				65%	4%				

~48W power saving with BLOB-DB workload analysis







Power Consumption for BLOB-DB



Power Provisioning through Trace-Driven Approach



Our Proposed Methodology

-   Disk power characterization by exercising synthetic I/O patterns
-   Analyze production workloads I/O-request traces and generate corresponding I/O profiles
-   Estimate power provisioning value for datacenter capacity planning using SPECpower + I/O profiles

Generating an App-Specific I/O Profile

Based on workload trace analysis, use IOmeter to tune ...

- Interarrival time
- Rnd/Seq ratio
- Read/Write ratio
- Block sizes

Edit Access Specification

Name: BLOB-DB | Default Assignment: None

Size	% Access	% Read	% Random	Delay	Burst	Alignment	Reply
OMB 8KB OB	60	100	93	0	1	sector	none
OMB 8KB OB	21	0	77	0	1	sector	none
OMB 64KB OB	4	100	93	0	1	sector	none
OMB 0KB 512B	6	0	77	0	1	sector	none
OMB 16KB OB	3	100	93	0	1	sector	none
OMB 1KB OB	6	0	77	0	1	sector	none

Transfer Request Size: 0 Megabytes | 8 Kilobytes | 0 Bytes

Percent of Access Specification: 60 Percent

Percent Read/Write Distribution: 0% Write | 100% Read

Percent Random/Sequential Distribution: 7% Sequential | 93% Random

Burstiness: Transfer Delay 0 ms | Burst Length 1 I/Os

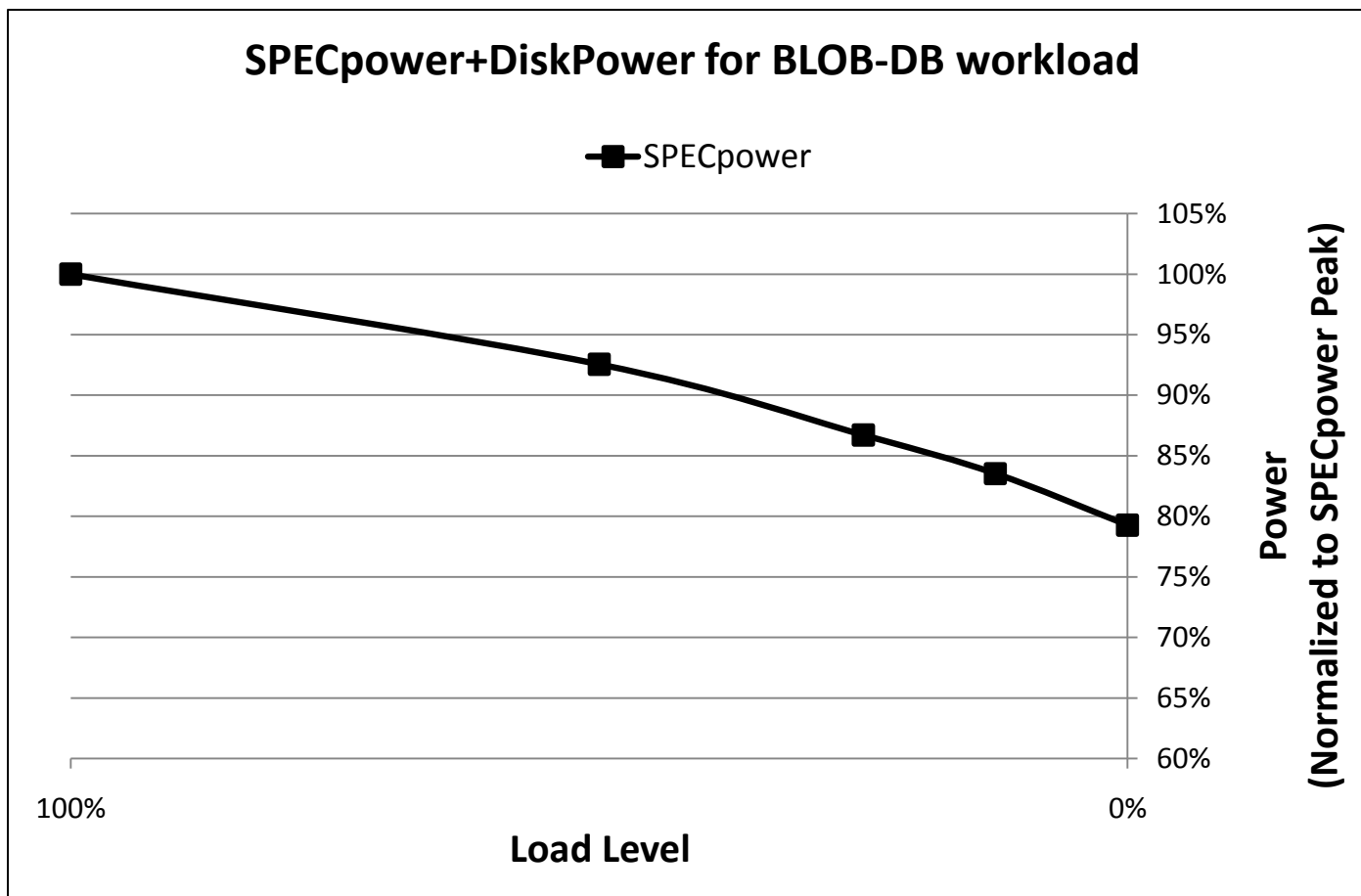
Align I/Os on: Sector Boundaries | 0 Megabytes | 0 Kilobytes | 512 Bytes

Reply Size: No Reply | 0 Megabytes | 8 Kilobytes | 0 Bytes

Buttons: Insert Before, Insert After, Delete, OK, Cancel

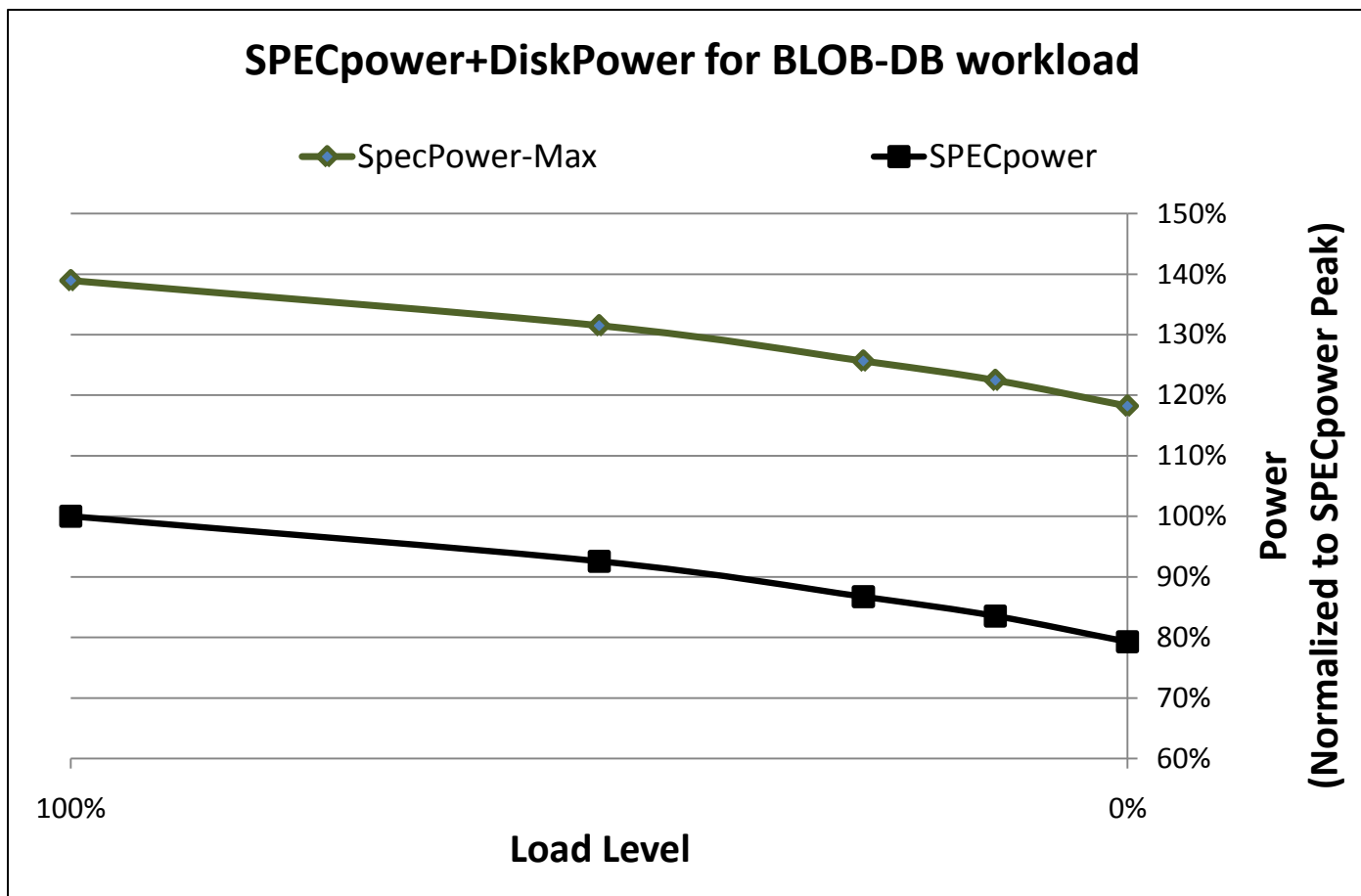
SPECpower + Disk Power (BLOB-DB)

SpecPower curve below does not include storage subsystem



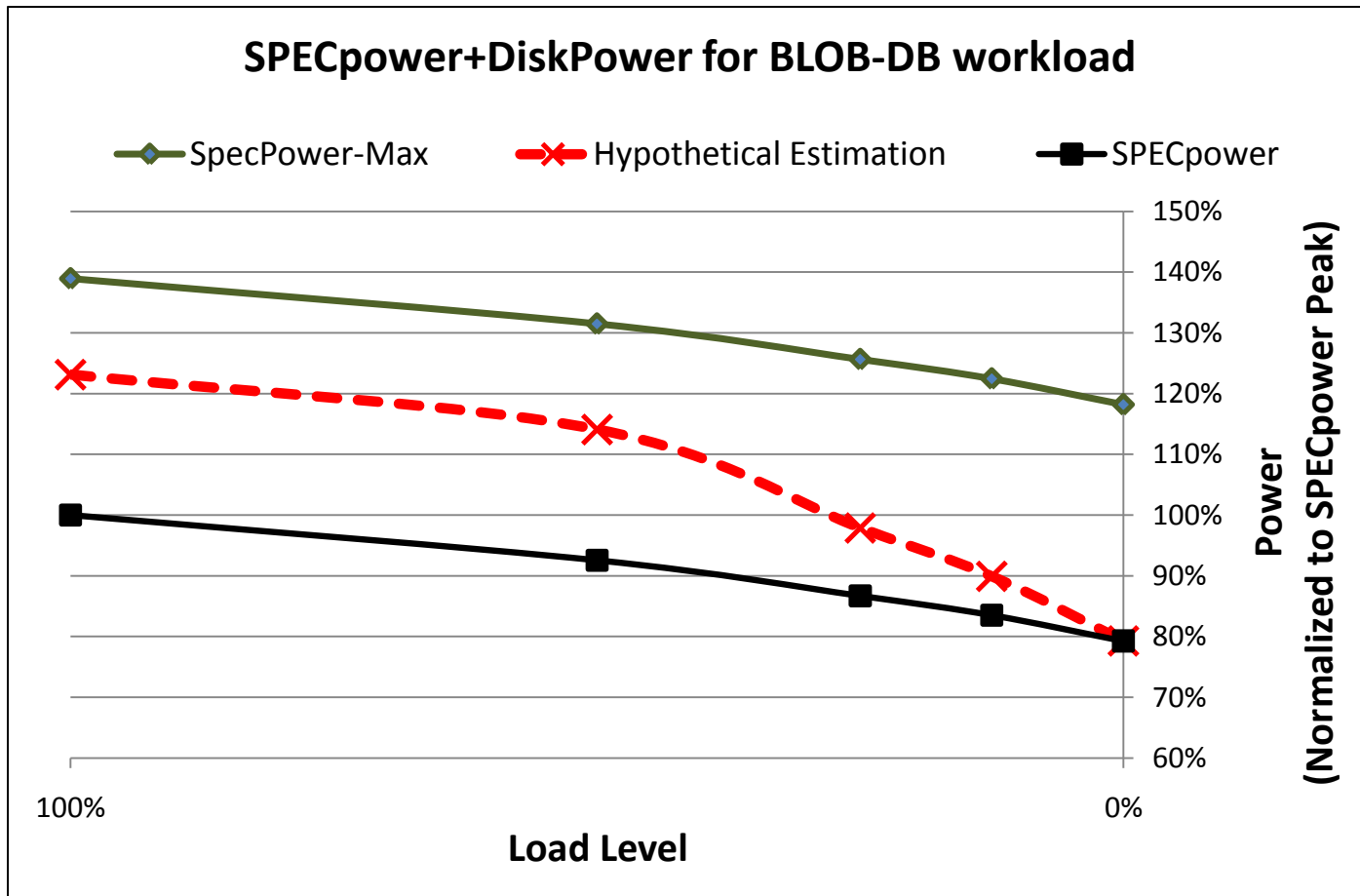
SPECpower + Disk Power (BLOB-DB)

SpecPower-Max assumes 100% 8K random load on all 37 drives in BLOB-DB



SPECpower + Disk Power (BLOB-DB)

Estimated values assumes workload specific profiles applied to storage subsystem



Cost Advantage of our approach





Datacenter Provisioning for BLOB-DB

Variables	Max-Power	Trace-Driven
Cost of Facility (\$):	\$200,000,000.00	\$200,000,000.00
Cost/Server (\$)	\$5,000.00	\$5,000.00
Size of Facility (Critical Load W):	15000000	15000000
Power/Server (W)	522	463
Number of Servers:	28735	32397
Scale units	1368	1542
Users on 1 unit	5000	5000
Total Users hosted	6.8M	7.7M

-900,000 more users in same datacenter

-Efficient Utilization of existing infrastructure

Contributions of the Paper

-  Detailed disk power characterization and application to actual production workloads
-  Generating application-specific I/O access profiles using trace-based workload analysis
-  Methodology to run SPECpower+IOmeter with specific app-profiles for determining server power
-  Using these power values for safely maximizing Server Capacity in given Datacenter power budget – significant cost savings



THANK YOU!

Q&A