ENERGY STAR
Data Center Server Meeting
Initial Insights from SERT
Server Results

John Clinger – ICF International

Al Thomason – TBWC, LLC

Agenda

1. High Level Observations and Notes on Dataset
2. Observations on Idle State Data
3. Observations on Active State Data
   - CPU Results
   - Larger Configurations
   - Memory Results
   - Storage Results
4. Conclusions, Open Discussion

Agenda

1. High Level Observations and Notes on Dataset
2. Observations on Idle State Data
3. Observations on Active State Data
   - CPU Results
   - Larger Configurations
   - Memory Results
   - Storage Results
4. Conclusions, Open Discussion

Initial High Level Observations

- First look at SERT based Active and Idle results
- Key goals:
  - Are there any readily noticeable show stoppers?
  - Investigate concerns surrounding large configurations
  - Identify trends in CPU, memory, and storage scores
  - Develop preliminary ideas for Version 3.0 approach

Initial High Level Observations

- No critical show-stopper noted
- Data supports separate categories
  - Rack, Blades, Resilient
- Storage is challenging
- Blades are efficient
- Memory requires additional segmentation
- Larger configurations test points are not clearly a problem

Observations from Dataset

- Some confusion over configuration points on 5-corner testing, examples:
  - Max Power configuration does NOT consume greatest power
  - Only 3 configuration points submitted

- Continued evidence of 'Idle Padding'
  - MUCH less what was observed in Version 1.0
    (with *massively* generous memory adder)
  - Examples:
    - Min system sold with 32GB; min tested 196GB
    - Min system sold with 128GB; min tested 256GB
    - 8GB "Low-end performance" vs. 64GB "Lowest Power"
      (*Low-end Perf.* consumed 9w *LESS* power @ idle then *Lowest Power* configuration)

Difficulties in Assessing Dataset

- Errors in data entry, semi-automation should be considered
- Small sample sizes – especially in 4-socket servers
- Necessary at times to review detailed SERT result sheet to:
  - Identify type of HDDs used, not just quantity.
  - Determine presence of other added cards, RAID controllers, NICs, etc.
  - Validate configuration details
  - Draw Insights from raw performance vs. power consumption (early assessments)
- Definition of test points and 'idle padding'
  - Must keep in mind V2 goals when assessing dataset– likely different then V3 goals.
- Unexplored / explained / unexpected results, e.g.:
  - ID #415 (High Performance Config) Raw SSJ results 1/4th of ID #461 (Typical Config).
- Inconsistencies between EPA SERT results and ITI spreadsheet e.g.:
  - ID #415 SSJ results: ITI=19.9, EPA SERT = 19.5
  - ID #435 SSJ results: ITI=37.5, EPA SERT = 15.6

Recommendation: Education Needed!

- Certification Bodies: Better review of submitted data to ensure:
  - Submission contains all data test points
  - Configurations appear to meet definitions for each test configuration

- Venders: Realize impact of submitted system test points
  - 5-point testing is intended in part to define an envelope of tested and certified systems
  - Submitting test points using 128GB and greater of memory does not support sale of 32GB systems under ENERGY STAR label

Other Recommendations:

- Industry should work to develop white-papers which explain:
  - How to work with the SERT results
  - What the results mean

- Start thinking about Version 3 changes:
  - Different guidance then current 5-corner box?

Dataset in Summary

- Quantity of configurations :
  - 4-Socket Resilient Servers: 30 - representing 3x systems
  - 2-Socket Resilient Servers: 18 - representing 2x systems
  - 4-Socket Managed Servers: 15 - representing 1x blades and 2x rack/tower systems
  - 2-Socket Managed Servers: 118 – representing 9x blades and 16x rack/tower systems
  - 1-Socket Managed Servers: 64 – representing 13x rack/tower systems

Note: From data pool dated March 25, 2014

May have been slightly updated in later releases.

Agenda

1. High Level Observations and Notes on Dataset
2. **Observations on Idle State Data**
3. Observations on Active State Data
   - CPU Results
   - Larger Configurations
   - Memory Results
   - Storage Results
4. Conclusions, Open Discussion

Idle Measurement & Margin

Median	High-End Performance	Low End Performance	Typical	Maximum Power	Minimum Power
2 Socket	44%	55%	58%	42%	43%
1 Socket	53%	43%	51%	46%	40%
1 Socket (Unmanaged)	51%	59%	56%	66%	51%

- Qualifying (V2) systems easily met Idle regiment at all test points

- Max % show worst case idle margin.
  - Still sufficient margin

Max	High-End Performance	Low End Performance	Typical	Maximum Power	Minimum Power
2 Socket	80%	69%	80%	84%	62%
1 Socket	61%	69%	69%	56%	53%
1 Socket (Unmanaged)	61%	69%	69%	56%	53%

% shown represent median of systems measured power draw relative to allowed.

Idle - Observations

- Observation:
  - Certified systems easily meet current Idle requirements
  - Keep in mind when deciding what to use for V3
    - SERT tool easily allows for Idle measurement
    - Idle in ENERGY STAR has a long history, customers and venders are comfortable with it

Agenda

1. High Level Observations and Notes on Dataset
2. Observations on Idle State Data
3. Observations on Active State Data
   - CPU Results
   - Larger Configurations
   - Memory Results
   - Storage Results
4. Conclusions, Open Discussion

Summary of Categories

- Median SERT values by classification

	CPU							Memory		Storage			Power	
Server	Compress	CryptoAES	LU	SOR	XML Validate	SORT	SHA256	Flood	Capacity	Sequential	Random	SPECPower SSJ Hybrid	Maximum Power	Idle Power
4S - Resilient Median:	**20**	**15**	**19**	**26**	**20**	**30**	**19**	**336**	**1114**	**23**	**7**	**27**	**1156**	**752**
4S - Blade Median:	**34**	**27**	**27**	**38**	**26**	**34**	**25**	**719**	**1628**	**94**	**68**	**36**	**3188**	**766**
4S - Rack Median:	**28**	**113**	**24**	**22**	**22**	**31**	**29**	**183**	**665**	**46**	**22**	**38**	**718**	**221**
%CH - Rack & Blade	-18%	319%	-11%	-41%	-14%	-9%	18%	-75%	-59%	-51%	-68%	3%	-77%	-71%
2S - Resilient Median:	**16**	**12**	**15**	**21**	**16**	**23**	**15**	**288**	**760**	**50**	**17**	**21**	**591**	**300**
2S - Blade Median:	**39**	**32**	**31**	**43**	**30**	**38**	**29**	**107**	**358**	**129**	**79**	**44**	**536**	**199**
2S - Rack Median:	**32**	**30**	**30**	**41**	**27**	**36**	**32**	**89**	**180**	**83**	**63**	**34**	**307**	**134**
%CH - Rack & Blade	-18%	-4%	-4%	-6%	-11%	-5%	10%	-17%	-50%	-35%	-20%	-23%	-43%	-33%
1S - Rack Median:	**33**	**36**	**30**	**35**	**29**	**31**	**31**	**25**	**54**	**113**	**60**	**38**	**117**	**46**

- Shows support for continued segmentation
- Blade systems overall show greater energy efficiency then rack systems

Note: From data pool dated March 25, 2014

Clustering of Active Results

- Percent change median vs. mean SERT results

	CPU							Memory		Storage		SSJ	Max Watts	Idle Watts
4S-Rack	0%	-1%	3%	0%	1%	1%	1%	-21%	-20%	-84%	-499%	-1%	-3%	-10%
2S-Rack	-6%	-29%	-3%	2%	-8%	-2%	-5%	-77%	-248%	-136%	-393%	-9%	-8%	-12%
1S-Rack	-4%	-79%	-11%	-9%	-6%	-9%	-2%	-13%	-67%	-70%	-143%	-4%	-7%	-21%
4S-Blade	-5%	-7%	-7%	-10%	-6%	-8%	-10%	26%	-2%	-25%	-300%	6%	17%	-16%
2S-Blade	-1%	-8%	-8%	-1%	-9%	-2%	-15%	-60%	-78%	-10%	-379%	-3%	-80%	-48%
4S-Resil.	1%	-4%	0%	-3%	0%	-1%	-1%	-54%	-4402%	-9%	-242%	2%	-4%	-7%
2S-Resil.	-5%	-5%	-5%	-4%	-5%	-7%	-5%	-12%	-21%	-10%	-244%	-5%	0%	-7%

- Consistency around CPU results
- Wide range of memory and storage results
  - Indicative of large variations of configurations supported by some systems

Note: From data pool dated March 25, 2014

Data Summary Observations

- Pool size is limited in some groupings
  - Specifically 4S servers
- Shows support for continuation of current segmentation
- Good clusters around CPU results
- Memory and Disk worklets show wide variations
  - Indicative of wide configuration ability of some machines
  - Will likely need additional work for V3 certification levels

Agenda

1. High Level Observations and Notes on Dataset
2. Observations on Idle State Data
3. Observations on Active State Data
   - CPU Results
   - Larger Configurations
   - Memory Results
   - Storage Results
4. Conclusions, Open Discussion

Larger Configuration Points vs. CPU Results

- In general - efficiency increases in “High Performance” configurations (vs. Typical configurations)
- Resilient Server data is exception.
  - Little change Max Power vs. High Performance
  - Both in results *and* configuration details.

Median SSJ results	SSJ Work/Watt % Change (Typical vs.)		Delta
	Max Performance	Max Power	
Resilient	(14%)	(16%)	2%
4 Socket	20%	(6%)	15%
2 Socket	2%	(13%)	11%
1 Socket	5%	(3%)	20%
Unmanaged	23%	(7%)	39%

Large Configuration Details

- Example of specific Resilient machines:

2 Socket:

#13	– Family (q)	High Perf	SSJ:	23	(Raw:	20.5	/	710w)	--	512GB,	6x-SSD
#1	– Family (q)	Typical	SSJ:	21	(Raw:	17.7	/	536w)	--	256GB,	3x-HDD
#31	– Family (r)	High Perf	SSJ:	15	(Raw:	20.9	/	894w)	--	1024GB,	8x-SSD
#33	– Family (r)	Typical	SSJ:	19	(Raw:	17.6	/	566w)	--	256GB,	3x-HDD

4 Socket:

#57	– Family (u)	High Perf	SSJ:	21	(Raw:	52.3	/	1,553w)	--	1024GB,	6x-SSD
#61	– Family (u)	Typical	SSJ:	30	(Raw:	53.3	/	1,119w)	--	256GB,	2x-HDD

- Notes
  - R & U –larger packaging with much greater expansion capability
  - SSJ workload does not take into account additional I/O capability
  - However, performance/watt results are impacted by additional power draw

- Example of ENERGY STAR working?
  - Median for SSJ resilient is 21. U & Q meet this, R does not.
  - Less energy efficient machines not awarded label?

Large Configuration Details

- Example of specific non-resilient machines:

2 Socket:

#91	– Family (m)	High Perf	SSJ:	36	(Raw:	100.7	/	2,247w)	--	512GB,	2x-HDD
#204	– Family (m)	Typical	SSJ:	45	(Raw:	74.2	/	1,315w)	--	32GB,	2x-HDD
#122	– Family (x)	High Perf	SSJ:	43	(Raw:	23.6	/	431w)	--	256GB,	8x-SSD
#318	– Family (x)	Typical	SSJ:	45	(Raw:	19.3	/	298w)	--	64GB,	4x-HDD

4 Socket:

#415	– Family (i)	High Perf	SSJ:	20	(Raw:	43.3	/	2,117w)	--	515GB,	4x-SSD
#461	– Family (i)	Typical	SSJ:	37	(Raw:	150.1	/	3,199w)	--	512GB,	4x-HDD

- Family ‘x’ showed 13% gain in SSJ results
  - 22% increase in RAW
- Family ‘m’ and ‘i’ showed decline in SSJ results
  - Question why such large RAW drop on ‘i’
- All three systems returned good results across all workloads
  - Largely all above median values
  - Despite differences in work/watt and RAW, indication each could ‘qualify’ in V3

Large Configuration Details

- Comparing RAW SSJ results between Typical and High Performance
  - All systems showed expected increase in energy draw
  - Many machines showed expected increase in RAW performance
    - However did not always outpace like increase in energy draw.
  - Some did not.
    - One showing 2/3 decline in RAW results – perhaps error in testing, data entry?
  - Results also impacted by:
    - Large I/O capability included in many examples.
    - Limited CPU variation in product line.
- Even so, some system showed overall good results – with overall improved work/watt in higher configurations.
- Large resilient Servers will need more detailed investigation
  - Understand configuration points, especially I/O content
  - Understand unexpected results, ala declined RAW performance.

Other Specific Observations

(non-resilient servers showing decline perf/watt: typical vs. high perf config)

- #415 4-socket server
  - Dramatic 73% *reduction* in raw SSJ results
  - Like reduction NOT seen in other CPU worklets
  - Should be investigated – seems like error
- #102 2-socket blade server
  - Little gain in raw SSJ results – used same CPU in typical and high performance config.
  - Chassis contained 20x HDDs in High performance config
    - Power consumption up greatly, CPU SSJ results basically the same...
- #91 2-socket blade
  - Raw SSJ results increased only 25%
  - Power consumption increased 100%
  - Other Active raw results along same line of increase
  - Slight increase in I/O
  - Massive increase in memory

    (128GB -> 2TB installed in system)

Comments -- Large System Configurations

- Concerned resilient data pool containing heavy configurations at High Performance
  - Why so similar to Max Power configurations?
  - Is this skewing sample set?
  - Are such heavy configurations really appropriate for High Performance?
    - Perhaps it is. . . .
- Most results show mixed bag
  - SERT results give indication large configurations at great disadvantage
  - Details do not always support same
- Even if SERT results do decline, several examples of machines likely to still 'certify' under V3
- Other systems producing poor results masquerade behind 'large configuration' issue
  - Poor results to begin with -- despite greater energy usage
  - Do not scale much -- despite additional resources

Larger Storage vs. CPU Results

- In data pool:
  - 15 configurations had > 8x HDDs (5x families)
  - 1 exceeded most-all 'median' CPU workload scores
    - 3x did well at other test points in family (with smaller # of HDDs)
    - 8x did poorly at all configuration test points.
  - 66% of large HDD configurations showed good storage results.
    - 33% showed poor perf/watt storage workload results

- Finer Focus
  - Reviewing only High Performance, Low Performance and Typical Configurations
    - 2x families are marginal: Doing well in 1 of 3 configurations
      - One system only did well in large HDD configuration! (#428)
    - 3x families did poorly in all configurations
    - No families did well at all configuration points.

Example of 'Marginal' Families
(rr & ee)

Unique ITI Identifier	Family	Server Type	Configuration	Processor Name	Total GB of mem.	Number of HDDs	HDD Speed	Compress	Crypto AES	LU	SOR	XML Validate	SORT	SHA256	Flood	Capacity	Sequential	Random	SPECPower SSJ Hybrid	Maximum Power	Idle Power	
87	ee	Managed	High-End Performance	E5-2650 V2	1536	16	15 K	15	21	22	29	18	26	24	12	86	21	9	15	531	313	
83	ee	Managed	Low End Performance	E5-2640 V2	16	1	7.5 K	47	42	41	58	38	52	47	23	91	14	8	49	219	101	
85	ee	Managed	Typical	E5-2640 V2	32	3	15 K	24	28	29	40	25	35	32	20	126	40	16	25	305	170	
428	rr	Managed	High-End Performance	E5-2450 v2	192	16	15 K	37	150	41	42	28	37	34	275	988	459	207	41	417	174	
432	rr	Managed	Low End Performance	E5-2403 0	16	2	7.5 K	32	26	26	28	26	24	22	36	58	35	22	34	115	62	
438	rr	Managed	Typical	E5-2407 v2	48	4	15 K	29	99	32	23	23	20	23	120	180	178	93	33	192	109	
					64	2		36	31	30	42	29	37	30	101	254	108	65	40	355	147	Median

- Family 'ee' did well only at low-end configuration point
- Family 'rr' did well at high performance configuration point
  - Large HDD configuration
- 'rr' has twice the disk capacity (16x vs 8x), slightly larger I/O (6x-pci vs 4x), half the DIMM capability.
  - 'ee' gives impression large configurations disadvantaged
  - 'rr' gives opposite impression
- Details give slightly different insight.

Details of 'Marginal' Families
(rr & ee)

		Family ee				Family rr		
		Typical	High Perf	% Change		Typical	High Perf	% Change
---	---	---	---	---	---	---	---	---
CPU	Compress	9.1	9.5	5%		7.0	18.4	162%
	CryptoAES	10.9	14.1	29%		24.0	76.9	221%
	LU	11.2	14.7	32%		7.5	19.6	163%
	SOR	14.2	18.5	30%		5.2	17.9	245%
	XMLvalidate	10.0	12.7	27%		5.6	13.4	138%
	Sort	13.1	17.1	31%		4.6	16.5	261%
	SHA256	11.8	15.5	31%		5.3	14.4	174%
Memory	Flood	7.5	7.8	5%		30.6	146.7	379%
	Capacity	57.9	73.4	27%		39.1	453.7	1061%
Storage	Sequential	9.1	9.0	-1%		29.3	144.8	394%
	Random	3.6	3.7	1%		14.9	61.1	311%
Hybrid	SSJ	9.8	10.4	6%		8.7	21.8	150%

- Family 'ee'
  - Produced good RAW results
  - Results did *not* scale well
    - Average 24% increase in computation oriented workloads
  - Storage workloads remained flat or declined
- Family 'rr'
  - Produced good RAW results
  - Results did scale well
    - Average 189% increase in computation oriented workloads
  - Storage workloads increased as well.

- Family 'ee'
  - Design does not scale well
  - Disk subsystem produced very poor results
  - Also consumes more power
  - These combined are perhaps *true* reason for poor results at larger configuration points.
- Family 'rr'
  - Produced good results
  - Scales well
  - Likely other configuration points could be improved with tweaking?

Review of RAW Results:

- Family ‘ee’
  - Design produced good results, but does not scale well
  - Disk subsystem produced very poor results
  - Also consumes more power
  - These combined are perhaps *true* reason for poor results at larger configuration points.

- Family ‘rr’
  - Produced good results
  - Scales well
  - Likely other configuration points could be improved with tweaking?

- Are these reflective of the systems architecture?
  - Or the configurations selected to meet V2 ENERGY STAR goals?

Recommendation -- Large System Configurations

- Better / refined guidance for configuration points
  - Especially around installed I/O capability
    - SERT has no I/O workloads to showcase / offset heavy I/O capabilities.
    - Perhaps even in resilient servers one would be advised to not configure I/O so heavy when doing CPU centric workloads - such as SERT
  - Consider consistent guidance for Storage in SUTs

- Continue to asses specific examples of large configurations
  - Do they contain excessive I/o capabilities?
  - Does underlying system show poor raw results?
  - Do other example machines show the ability to scale well?
    - Ala: family ‘rr’, ‘q’ and ‘x’

- If justified - investigate multipliers to CPU related workloads for:
  - Large memory configured machines?
  - Large HDD configured machines?

Agenda

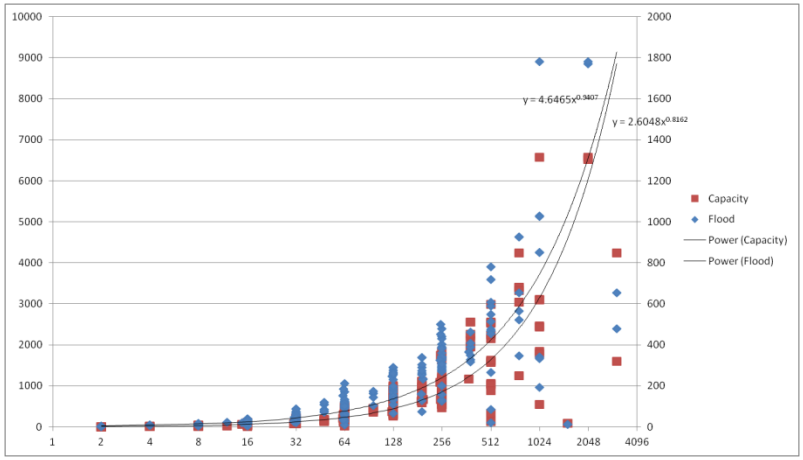
1. High Level Observations and Notes on Dataset
2. Observations on Idle State Data
3. Observations on Active State Data
   - CPU Results
   - Larger Configurations
   - Memory Results
   - Storage Results
4. Conclusions, Open Discussion

Memory in Servers – Dynamic

- Servers can contain small to massive amounts of memory
  - Best Practices call for no disk page-swapping
- Data pool ranged from 2GB to over 3TB RAM
- Single threshold likely not to suffice in V3

Linearity of Memory Tests

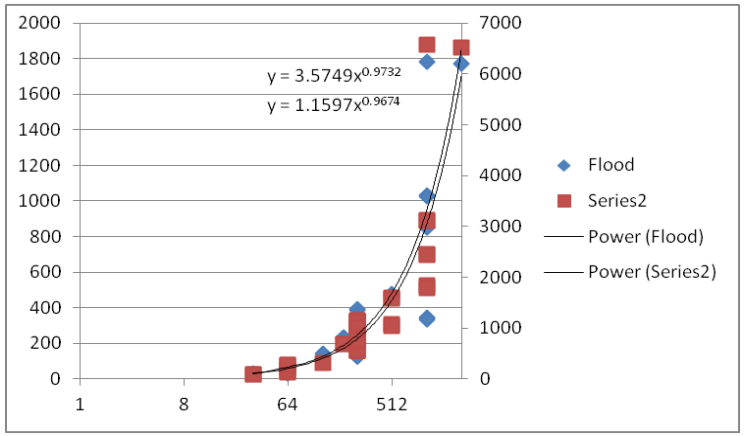
- Plotting all submitted systems (all configurations)
  - Memory Size vs. SERT results.
  - Using Excel ‘trend line’ to ‘fit’ data.
- Results rather liner
  - See X exponent clustering around 1.0



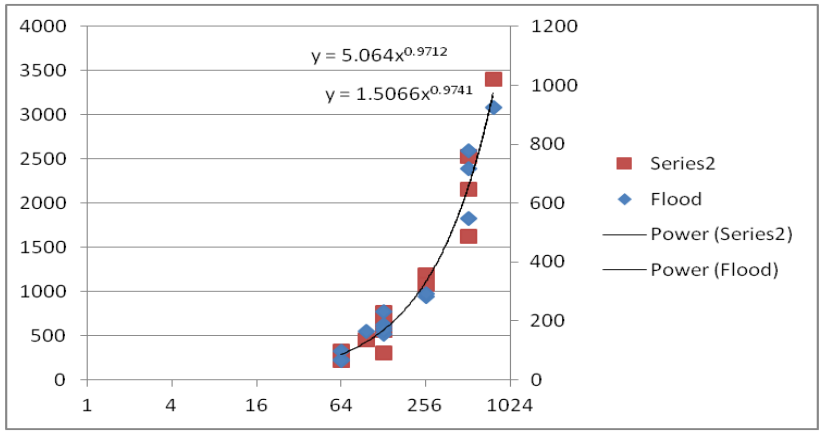
	Flood	Capacity
Resilient	$y = 1.1597x^{0.9674}$	$y = 3.5749x^{0.9732}$
4 Socket	$y = 1.5066x^{0.9741}$	$y = 5.064x^{0.9712}$
2 Socket	$y = 4.0032x^{0.7131}$	$y = 7.398x^{0.8416}$
1 Socket	$y = 2.6586x^{0.8163}$	$y = 3.4106x^{1.0726}$
Unmanaged	$y = 2.0562x^{0.8489}$	$y = 1.9161x^{1.2389}$
Average (All data)	$y = 2.6048x^{0.8162}$	$y = 4.6465x^{0.9407}$

All Systems Plot. Note use of Log(2) scale; hence curve look to ‘liner’ trend line.

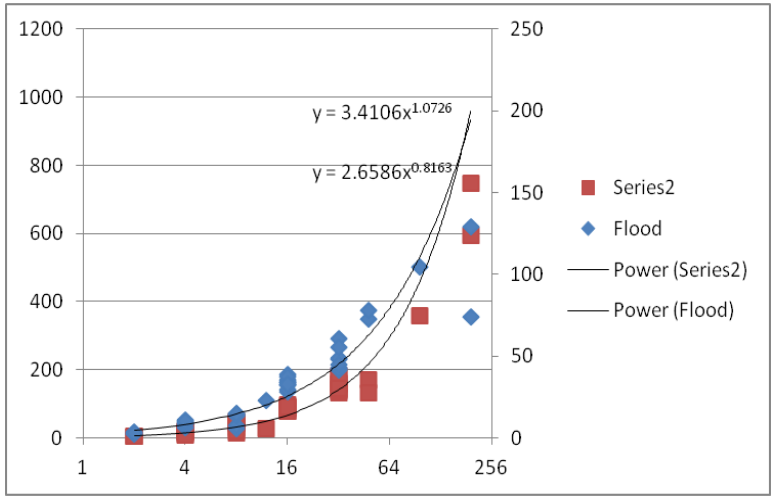
Additional Memory Plots



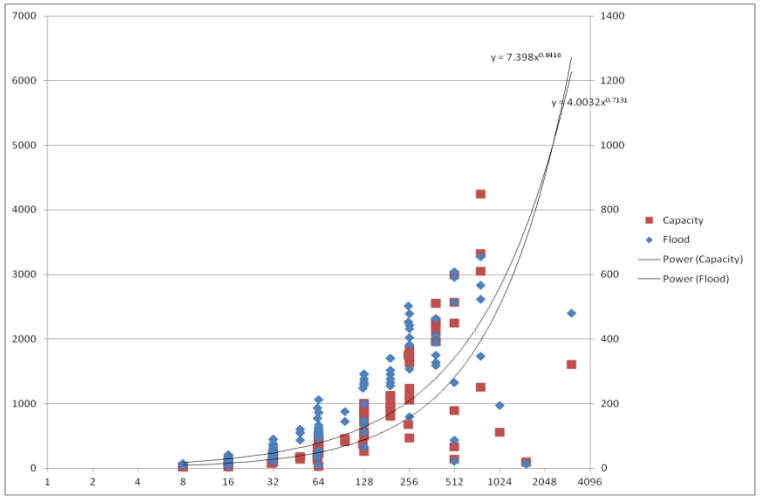
Resilient Server



4 Socket Server



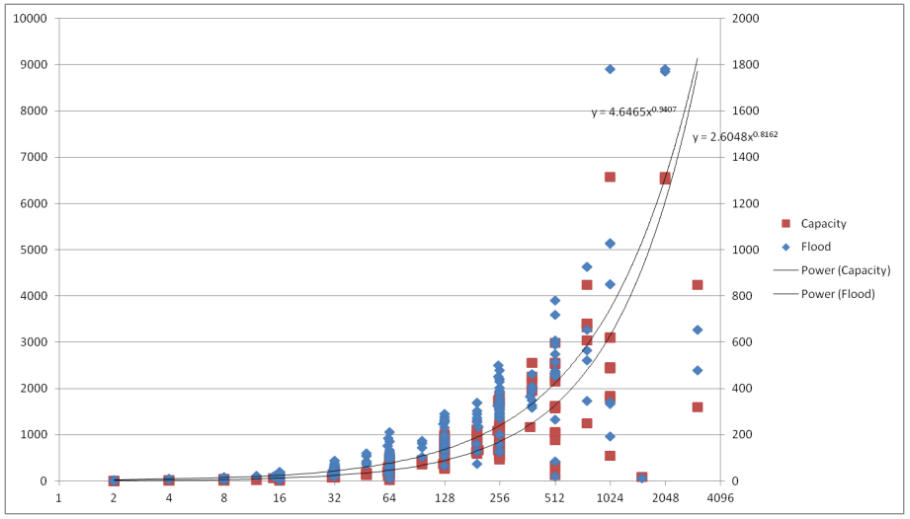
1 Socket Server

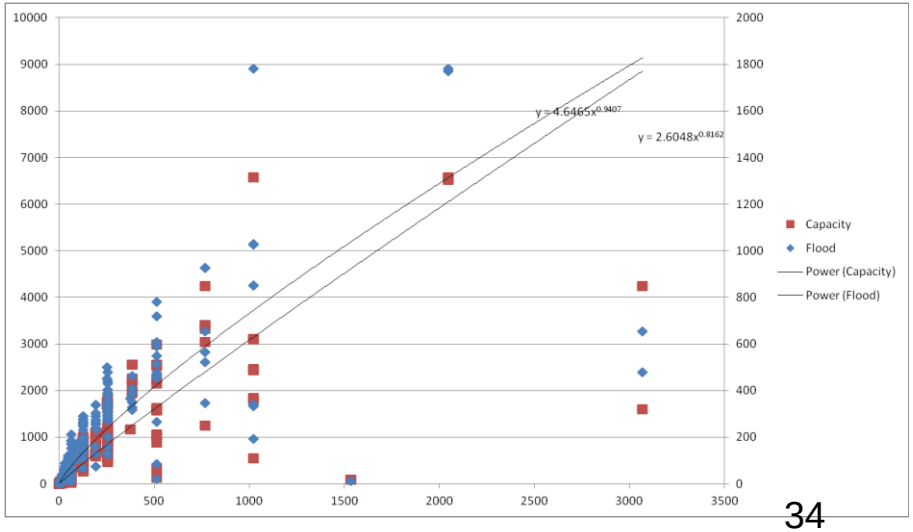


2 Socket Server

Side-track: Log vs. Linear Graphs

- Prior graphs show using log(2) as X scale
  - Prevents 'bunching up' of results at low end.
  - Provided clearer visual picture of trends
- Makes 'liner-ish' trend lines look like a curve…
- These two graphs are the same data
  - One in log(2) form.
  - Other Liner form.





Observations - Memory

- Data shows a surprising amount of linearity
  - SERT tests results vs. installed GB of memory
  - Each ‘class’ of servers have slightly different rate of changes
- Likely need different thresholds for each class of servers

Agenda

1. High Level Observations and Notes on Dataset
2. Observations on Idle State Data
3. Observations on Active State Data
   - CPU Results
   - Larger Configurations
   - Memory Results
   - Storage Results
4. Conclusions, Open Discussion

Storage in Servers - Complex

- Greatly complicating any active assessment of Storage is the wide usage models:
  - Boot / page-swap / cache only
  - Primary data storage (Shared or exclusive to local server applications)
  - No storage! (Boot from SAN)
- Each drives other architecture considerations:
  - Cooling, RAID cards, Power Supply, etc.
- Difficulty arrives with how additional storage devices impact work/watt results
  - CPU orientated workloads unlikely to benefit in performance
  - While system sees increased power consumption

Observations from Storage Data

- 90% of systems showed increase
  - Unmanaged 1S servers being exception
- MASSIVE range of result within a family
  - Over 3,000% median change typical vs. High Performance

Storage SERT results % Change
(Typical to Max Performance configuration)

	Sequential	Random
Resilient Servers	158%	2,386%
4S	496%	3,216%
2S	118%	166%
1S	51%	149%
Unmanaged	(34%)	8%

- Fair correlation between good CPU results and good disk results
  - Of 18x family's that appear to do well on CPU intensive results, all but 5 also do well on Disk SERT results
  - Though many only at High Performance configurations.
- System showing reduction in SERT results often also reduced # of HDDs
  - Side-effect of 'idle padding'?

Agenda

1. High Level Observations and Notes on Dataset
2. Observations on Idle State Data
3. Observations on Active State Data
   - CPU Results
   - Larger Configurations
   - Memory Results
   - Storage Results
4. Conclusions, Open Discussion

Other Questions to be Addressed

- Adjust testing points
  - 5 corner box goes to three?
    - Lower testing cost eliminating highly configured (Max Power) test point.
    - Do end consumers see value in Max Power? (ala, rack power distribution planning??)

- Comment on ‘Hybrid Algorithm’
  - Values do not seem to track with proposed approach
  - Some smaller hybrid value systems and some larger ones meet above, others do not

Open Discussion

- Any questions or comments?

Contacts

- RJ Meyers, EPA
  - 202-343-9923
  - Meyers.Robert@epa.gov
- John Clinger, ICF International
  - 215-967-9407
  - John.Clinger@icfi.com
- Al Thomason, TBWC, LLC
  - 503-708-7881
  - thomasonw@gmail.com