Performance Analysis "New" Technology

Velocity Software Inc.
196-D Castro Street
Mountain View CA 94041
650-964-8867

Velocity Software GmbH
Max-Joseph-Str. 5
D-68167 Mannheim
Germany
+49 (0)621 373844

Barton Robinson,
barton@velocitysoftware.com

New Performance Management Technology Topics

Technology does not stand still – Performance Questions

- Linux, VSE,
- SMT Value
- HiperDispatch Value, affinity value
- Diagnose rates
- Z14 Capacity vs Z13
- FCP / EDEV Performance
- SPECEX impact
- GPFS Performance
- Extended Address Volume Minidisk support (Size of disk)
- More than 64 logical processors – impact? Performance?
- Diagnose type/rate by virtual machine (ESAUSRD)
- Docker

VELOCITY SOFTWARE

Performance Management Data Sources

Instrumentation sources

- NO Control Blocks (HIGH OVERHEAD)
- Standard, Defined APIs!!!
- No release to release issues

CP Monitor (VERY LOW OVERHEAD)

- Continuous enhancements

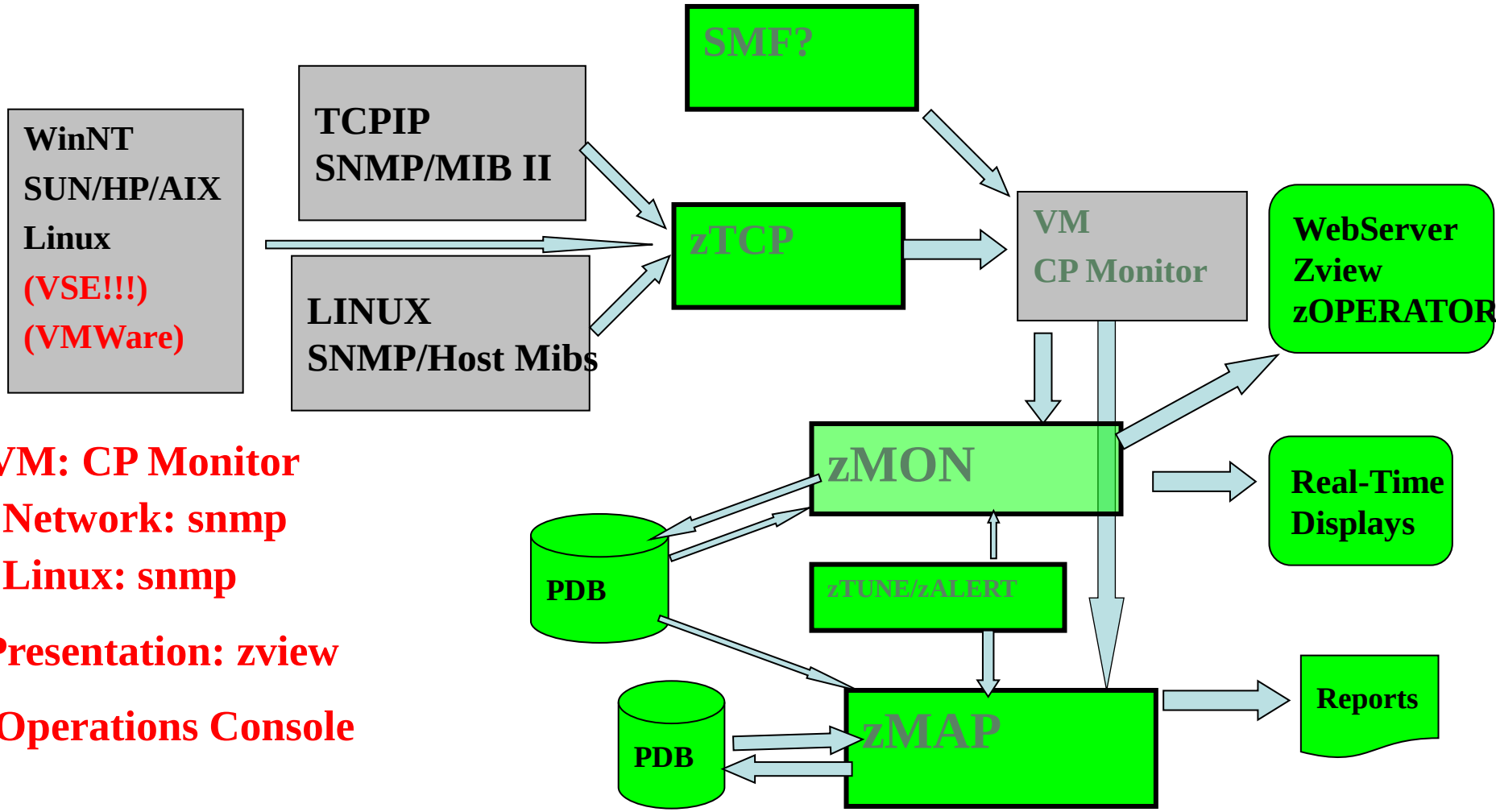
SNMP (VERY LOW OVERHEAD)

- Standard across ALL platforms!
- Easily enhanced
  - Velocity mib
  - UCD (Linux)
  - GPFS
  - Java / Websphere
  - VSE

SMF?

VELOCITY SOFTWARE

Velocity Integrated Architecture for 30 Years!!!



VM: CP Monitor

Network: snmp

Linux: snmp

Presentation: zview

Operations Console

VELOCITY SOFTWARE

New Performance Management Technology Topics

Data Available, not really understood?

- MFC – critical to evaluating “performance improvements”
- SMT – how much more capacity? Really?

“z/VM Continuous Delivery News” new stuff…

(some monitor changes are needed – please see your vendor)….

- zHPM CPU Resource Management (???)
- FCP Monitor Enhancements (ESAFCP/ESAEDEV)
- Extended Address Volume Minidisk support (ESASEEK)
- Cylinder / block max size (ds8880+)
- More than 64 logical processors (96 supported)
- Diagnose type/rate by virtual machine (ESAUSRD)

Linux “continuous delivery”…

- GPFS Support (ESAGPFSx)
- Linux diagnose table (ESALNXG)

VELOCITY SOFTWARE

Current Capacity Questions?

SMT: How Good or How Bad? It depends….

Depends on what?

- Are there cycles to spare?
- Is the cache thrashing?
- Is the TLB effective?

How much extra capacity on z14 vs z13?

- More cache?
- Better DAT?

What is impact of new patches? (specter)

- How to measure?

VELOCITY SOFTWARE

SMT Capacity Questions?

Z13/z14 have Multithreading (SMT-2)

- How much more (less) throughput?
- How to predict based on cycles lost for cache miss

Z13/z14 have larger processor cache

- Does affinity work to use the cache?
- How long does cache last when 50,000 dispatches / second / thread?

How much better is the z14? (REALLY BETTER…)

VELOCITY SOFTWARE

Why you should be interested if MFC

```
Report: ESAMFC          MainFrame Cache Analysis Rep
Monitor initialized: 12/23/14 at 13:55:31 on 2964
-------------------------------------------------
               <CPU Busy> <-------Processor------>
               <percent>  Speed/<-Rate/Sec->
Time      CPU  Totl User   Hertz Cycles Instr Ratio
--------  ---  ---- ----   ----- ------ ----- -----
14:05:32    0  92.9 64.6   5000M  4642M 1818M 2.554
            1  92.7 64.5   5000M  4630M 1817M 2.548
            2  93.0 64.7   5000M  4646M 1827M 2.544
            3  93.1 64.9   5000M  4654M 1831M 2.541
            4  92.9 64.8   5000M  4641M 1836M 2.528
            5  92.6 64.6   5000M  4630M 1826M 2.536
               ---- ----   ----- ------ ----- -----
System:         557  388   5000M  25.9G 10.2G 2.542
-------------------------------------------------
14:06:02    0  67.7 50.9   5000M  3389M 2052M 1.652
            1  67.8 51.4   5000M  3389M 2111M 1.605
            2  69.0 52.4   5000M  3450M 2150M 1.605
            3  67.2 50.6   5000M  3359M 2018M 1.664
            4  60.8 44.5   5000M  3042M 1625M 1.872
            5  70.1 53.8   5000M  3506M 2325M 1.508
               ---- ----   ----- ------ ----- -----
System:         403  304   5000M  18.8G 11.4G 1.640
```

1900 mips (at 100%)

2800 Mips (at 100%)

VELOCITY SOFTWARE

Processor speed maxed out

- Higher speeds “melt” chips
- Liquid cooled will be faster, just because they can

EC12 announcment:

- 2nd generation out of order design
- out-of-order superscalar chip
- branch prediction effectiveness
- A new set of instructions (requires compiler/user)
  - (Impacts real work done, not measureable in production)
  - (pauseless java for example)

Focus: Instructions / cycle

- Up to six instructions can be decoded per clock cycle.
- Up to ten instructions can be in execution per clock cycle.
- Instructions can be issued out-of-order.
- Memory accesses might not be in the same instruction order (out-of-order operand fetching).
- Several instructions can be in progress at any moment, subject to the maximum number of decodes and completions per cycle.
- SIMD: New set of instructions (requires compiler, user)
- SMT: allows two separate processes to run simultaneously on the same core.

Focus: Reduced cycles per Instructions

- SMT: allows two concurrent processes on one core.
- 1.5x more on-chip cache per core compared to the IBM z13™. Bigger (smaller??) and faster caches help to avoid untimely swaps and memory waits while…
- The new mechanism allows full out-of-order and branch speculation for any instruction that hits in L1/L2 caches, and allows for out-of-order TLB-miss handling.
- z14 includes a new translation lookaside buffer (TLB2) design with four hardware-implemented translation engines that reduces latency when compared with one pico-coded engine on z13

Why is cache, TLB important?

What is required for instruction to execute?

- Data, instructions, TLB to ALL be in L1 Cache

Each Cache level slower

How effective is cache?

Value of cache – Relative Nest Intensity (RNI)

IBM RNI calculation analysis (no account for TLB)

- **zEC12**
  RNI=2.3x(0.4xL3P+1.2xL4LP+2.7xL4RP+8.2xMEMP)/100

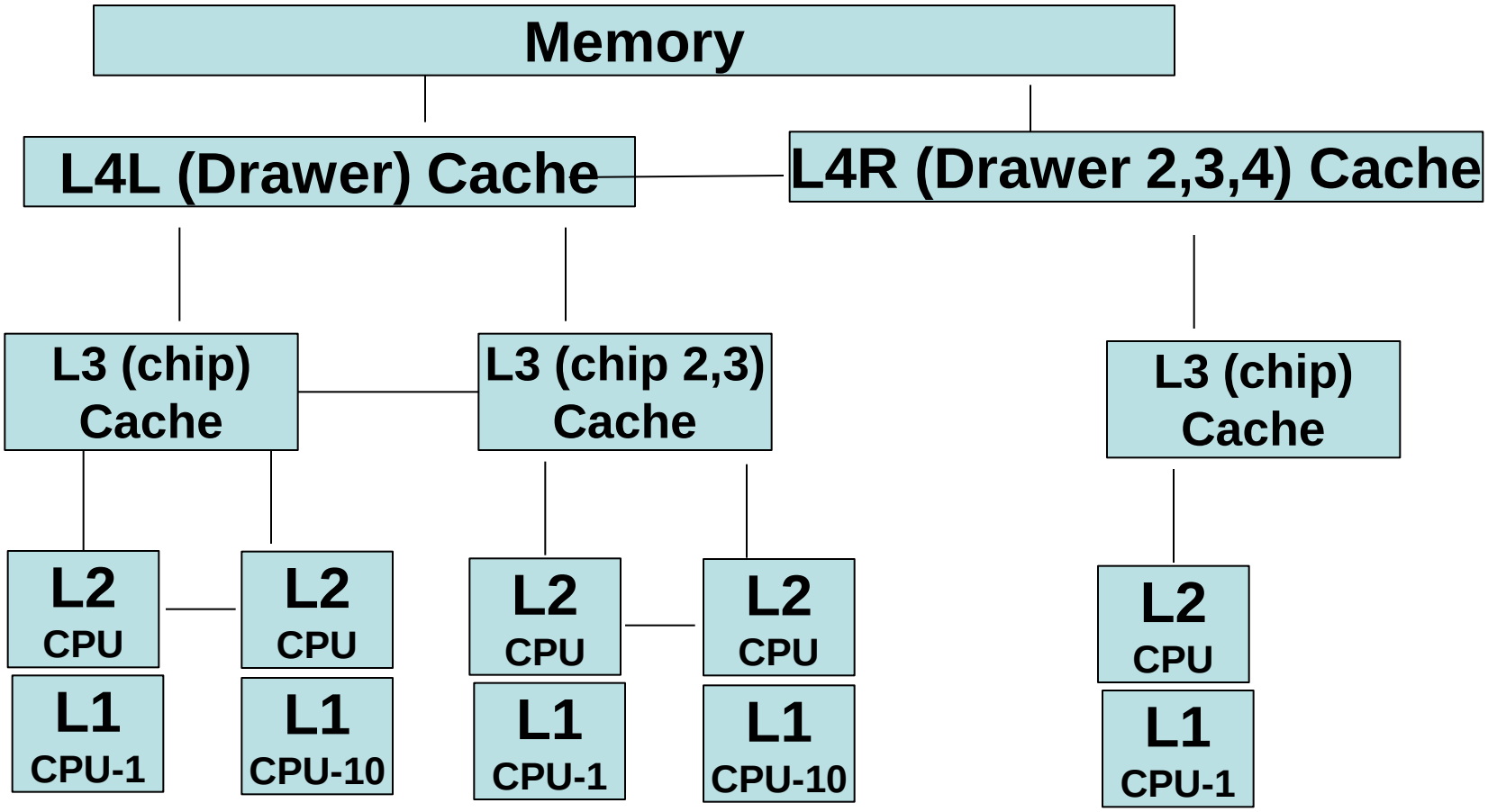
Cost analysis - ratio

- **L3P: 1 - L3 cache source**
- **L4LP: 3 - L4 local cache source**
- **L4RP: 6 - L4 Remote cache source**
- **MEM: 19 - memory source**

VELOCITY SOFTWARE

IBM RNI calculations (per John Burg)

- **z13**
  RNI=2.6x(0.4xL3P+1.6xL4LP+3.5xL4RP+7.5xMEMP)/100
- **zEC12**
  RNI=2.3x(0.4xL3P+1.2xL4LP+2.7xL4RP+8.2xMEMP)/100
- **z196**
  RNI=1.67x(0.4xL3P+1.0xL4LP+2.4xL4RP+7.5xMEMP)/100
- **z10**
  RNI=(1.0xL2LP+2.4xL2RP+7.5xMEMP)/100.



Question, If 50,000 dispatch / second / cpu, impact?

VELOCITY SOFTWARE

Processor cache comparison perspective

Cache Sizes – z13

- L1: 96K Instruction, 128K Data
- L2: 2MB Instruction, 2MB data
- L3: 64MB (Chip, Shared over 8 CPUS)
- L4: 480MB + 224M NIC (per node) (for 24 cpu)

Cache Sizes – Z14

- L1: 128K Instruction, 128K Data
- L2: 4 MB Data(includes L1), 2 MB Instruction(includes L1)
- L3: 128MB (over 10 cpus)
- L4: 672 MB (includes L3) (for 30 cpu)

Cache sizes marginally larger? One thread vs two?

VELOCITY SOFTWARE

Showing Value: CPU Measurement Facility

What is the CPU Measurement Facility

- Hardware instrumentation
- Statistics by LPAR, all guests aggregated
- 5.18 Monitor records (PRCMFC) (Basic, Extended)
- “Extended” different for z10, 196, EC12, z13 and z14
- Shows cycles used, instructions executed and thus CPI

```
Report: ESAMFC          MainFrame Cache Analysis Re
Monitor initialized: 02/27/15 at 20:00:00
-----------------------------------------------
                <CPU Busy> <-------Processor------>
                <percent>  Speed/<-Rate/Sec->
Time       CPU Totl User   Hertz Cycles Instr Ratio
-------- --- ---- ----   ----- ------ ----- -----
20:01:00    0  0.7  0.4   4196M  30.8M 8313K 3.709
```

VELOCITY SOFTWARE

CPU Measurement Facility (5.5Ghz = EC12)

What is the CPU Measurement Facility (Basic)

```
Report: ESAMFCA       MainFrame Cache Hit Analysi
Monitor initialized: 12/10/14 at 07:44:37 on 282
-----------------------------------------------
                <CPU Busy> <-------Processor------>
                <percent>  Speed/<-Rate/Sec-> CPI
Time      CPU  Totl User   Hertz Cycles Instr Ratio
-------- --- ---- ----   ----- ------ ----- -----
07:48:35   0 20.8 18.4   5504M  1121M  193M 5.807
           1 21.6 19.6   5504M  1161M  221M 5.264
           2 24.4 22.5   5504M  1300M  319M 4.078
           3 22.4 19.7   5504M  1248M  265M 4.711
           4 19.6 17.6   5504M  1102M  194M 5.683
           5 20.4 18.6   5504M  1144M  225M 5.087
           6 23.9 22.0   5504M  1341M  341M 3.935
           7 17.6 15.4   5504M   949M  160M 5.927
           8 18.5 16.5   5504M  1005M  194M 5.195
           9 22.5 20.6   5504M  1259M  347M 3.629
             ---- ----   ----- ------ ----- -----
System:       212  191   5504M  10.8G 2457M 4.733
```

VELOCITY SOFTWARE

What is the CPU Measurement Facility (Extended)

- L1 is cache misses. Small is good, “MEM” is expensive

```
Report: ESAMFCA     MainFrame Cache Hit Analysis
-------------------------------------------------
             <----Rate per 100 Instructions----->
             <-----Data source read from-------->
Time         L1     L2     L3     L4L    L4R    MEM
--------  -----  -----  -----  -----  -----  -----
07:48:35  3.605  2.062  0.948  0.247  0.003  0.346
          3.281  1.935  0.831  0.195  0.002  0.319
          2.607  1.656  0.577  0.137  0.001  0.237
          2.913  1.678  0.786  0.249  0.002  0.198
          3.572  1.973  1.037  0.330  0.002  0.230
          3.188  1.815  0.889  0.272  0.002  0.210
          2.410  1.462  0.605  0.187  0.002  0.156
          3.729  1.793  1.220  0.654  0.035  0.026
          3.209  1.593  1.017  0.535  0.029  0.036
          2.182  1.222  0.602  0.307  0.018  0.034
          -----  -----  -----  -----  -----  -----
System:   2.941  1.670  0.800  0.286  0.008  0.176
```

VELOCITY SOFTWARE

What to measure

- L1MP – Level 1 Miss %
- L2P – % sourced from L2 cache
- L3P – % sourced from Level 3 Local (chip) cache
- L4LP – % sourced from Level 4 Local book
- L4RP - % sourced from Level 4 Remote book

- MEMP – % sourced from Memory - EXPENSIVE

Why you should be interested – what is a MIP?

How to make CPI drop? Better (lower) L1 miss

Fix websphere, DB2….

```
Report: ESAMFCA    MainFrame Cache Hit Analysis  VelocitySoftware
-------------------------------------------------------------------------
             <CPU Busy>-----> <----Rate per 100 Instructions-----> 
             <percent>  CPI   <-----Data source read from-------->
Time     CPU Totl User  Ratio L1    L2    L3    L4L   L4R   MEM
-------- --- ---- ----  ----- ----- ----- ----- ----- ----- -----
14:05:32   0 92.9 64.6  2.554 4.618 3.963 0.585 0.042 0.000 0.023
           1 92.7 64.5  2.548 4.624 3.972 0.584 0.040 0.000 0.024
           2 93.0 64.7  2.544 4.587 3.928 0.590 0.042 0.000 0.023
           3 93.1 64.9  2.541 4.561 3.904 0.587 0.043 0.000 0.022
           4 92.9 64.8  2.528 4.542 3.888 0.585 0.042 0.000 0.023
           5 92.6 64.6  2.536 4.564 3.907 0.588 0.041 0.000 0.023
             ---- ----  ----- ----- ----- ----- ----- ----- -----
System:       557  388  2.542 4.582 3.927 0.587 0.042 0.000 0.023
-------------------------------------------------------------------------
14:06:02   0 67.7 50.9  1.652 2.456 2.115 0.302 0.020 0.000 0.016
           1 67.8 51.4  1.605 2.322 1.999 0.286 0.020 0.000 0.015
           2 69.0 52.4  1.605 2.273 1.945 0.290 0.023 0.000 0.013
           3 67.2 50.6  1.664 2.409 2.061 0.308 0.024 0.000 0.014
           4 60.8 44.5  1.872 2.952 2.535 0.371 0.027 0.000 0.017
           5 70.1 53.8  1.508 2.097 1.799 0.263 0.019 0.000 0.013
             ---- ----  ----- ----- ----- ----- ----- ----- -----
System:       403  304  1.640 2.391 2.052 0.300 0.022 0.000 0.015
```

VELOCITY SOFTWARE

TLB Analysis – z13 data SMT Enabled

Address Translation – the other single thread on z13

Why working sets are important,

Why we need large pages? (or better DAT?)

```
ESAMFC       MainFramate   ZMAP 4.2.2 08/10/15   Page  164
initialized: 07/07/157/07/15 13:04:00
-------------------------------------------------------------
     <CPU Busy> <---- <-Translation Lookaside buffer(TLB)->
     <percent>  Speed <cycles/Miss><Writs/Sec> CPU  Cycles
 CPU Totl User  Hertz  Instr Data  Instr Data   Cost  Lost
 --- ---- ----  -----   ----- ----- ----- ----- ----- -----
   0 26.4 24.2  5000M     102   534 1043K  527K  29.23  388M
   1 25.4 23.7  5000M     100   541 1010K  499K  29.12  371M
   2 24.5 22.8  5000M     127   558  872K  487K  31.09  383M
   3 25.8 24.1  5000M     125   554  891K  500K  30.06  389M
   4 20.0 18.3  5000M     131   575  667K  376K  30.19  303M
   5 21.1 19.6  5000M     126   579  679K  374K  28.53  302M
```

VELOCITY SOFTWARE

TLB Analysis – Should SMT be Enabled?

z/VM Linux workloads issue: VERY HIGH dispatch

Why z14 should be great….

Don't enable SMT if one thread is consuming your DAT

Report: ESAMFC MainFrame Cache Magnitudes Report

	<CPU Busy>		<----	<-Translation Lookaside buffer(TLB)->					
	<percent>		Speed	<cycles/Miss>		<Writs/Sec>		CPU	Cycles
##	Totl	User	Hertz	Instr	Data	Instr	Data	Cost	Lost
Mem1	907	874	5504M	54	232	117M	36M	29.55	14.8G
Mem2	1188	1140	5000M	147	364	30M	26M	23.62	14.0G
VLB4	1703	1366	5000M	185	567	66M	46M	44.59	38.2G
z13N	216	212	5000M	192	598	3084K	1802K	15.94	1669M
TCPN	892	757	5000M	217	947	32M	17M	51.46	23.0G
MTRN	947	868	5000M	265	1283	33M	17M	65.25	30.8G ←

VELOCITY SOFTWARE

TLB Analysis – Why is z/VM so bad?

z/VM Linux workloads issue: VERY HIGH dispatch

Validate data, 6 intervals, consistently bad

Report: ESAMFC MainFrame Cache Magnitudes Report

	<CPU Busy>		<----	<-Translation Lookaside buffer(TLB)->					
	<percent>		Speed	<cycles/Miss>		<Writs/Sec>		CPU	Cycles
##	Totl	User	Hertz	Instr	Data	Instr	Data	Cost	Lost
System:	947	868	5000M	265	1283	33M	17M	**65.25**	30.8G
System:	1001	963	5000M	200	1693	36M	13M	57.40	28.7G
System:	956	895	5000M	271	1314	32M	16M	62.91	30.0G
System:	954	874	5000M	272	1282	33M	17M	65.85	31.3G
System:	965	896	5000M	277	1306	33M	17M	66.02	31.8G
System:	945	870	5000M	273	1303	33M	17M	66.24	31.2G

VELOCITY SOFTWARE

TLB Analysis – Why is z/VM so bad?

z/VM Linux workloads issue: VERY HIGH dispatch

Validate data, 6 intervals, consistently bad

```
Report: ESAPLDV      Processor Local Dispatch                      ZMAP 4
Monitor initialized: 05/27/16 at 06:55:52 on cord analyzed: 05/27/16 06:5
-------------------------------------------------------------------------
         <----Load---->                                  <-CPU Steals fr
         <-Users-> Tran     <VMDBK Moves/sec>> Dispatcher <-From Nesting
Time     Actv In Q /sec CPU Steals  To Mastery Long Paths Same  NL1  NL2
-------- ---- ---- ---- --- ------  ---------- ---------- ---- ---- ----
06:57:00   97  266  0.6    0 9094.5        8.00    25011.3 6547 2548    0
                           1 9316.9          07    25456.0 7052 2265    0
                           2 9108.4          00    23481.9 7177 1932    0
….
                          10 7610.1          03    19939.7 6030 1580    0
                          11 7752.8          07    19363.9 6256 1497    0
                             --     -------    ------- ---- ---- ----
System:                     109310        8.07   274226.8  73K  37K    0
```

VELOCITY SOFTWARE

TLB Analysis – Why is z/VM so bad?

z/VM Linux workloads user level: VERY HIGH dispatch "RSCS" BUG

Report: **ESAUSR3** User Resource Utilisation - Part 2

DASD MDisk Virt <Dispatch> UserID /Class	DASD I/O	Block I/O	Cache Hits	Disk I/O	<Rate/Sec> Disp	Waits	
06:57:00	2787	9	1047	0	194K	194K	
***Key User Analysis ***							
RSCS	0	0	0	0	15K	15019	←-------
TCPIP	0	0	0	0	15K	14805	
User Class Analysis							
ZVPS	296	0	85	0	299	299	
TheUsers	2471	0	954	0	164K	164K	
Top User Analysis							
CV52D172	85	0	0	0	3095	3095	
CV52D157	60	0	0	0	3787	3787	
CV52D160	40	0	0	0	4684	4684	
CV52D154	29	0	0	0	8098	8098	
CV52D152	18	0	0	0	3218	3218	
CV52D156	31	0	0	0	3068	3068	

VELOCITY SOFTWARE

TLB Analysis – Why is z/VM so bad?

z/VM Linux workloads user level: VERY HIGH dispatch "WAS"…

Report: ESAUSR3 User Resource Utilisation - Part 2

DASD MDisk Virt <Dispatch> UserID /Class	DASD I/O	Block I/O	Cache Hits	Disk I/O	<Rate/Sec> Disp	Waits
15:03:00	15432	1272	2579	0	**120K**	**120K**
***Key User Analysis ***						
RSCS	1	0	0	0	5	5
TCPIP	0	0	0	0	250	250
User Class Analysis						
Servers	7203	0	2347	0	95	95
ZVPS	286	0	100	0	207	207
TheUsers	7942	1272	132	0	120K	120K
Top User Analysis						
CV52D027	16	0	0	0	1440	1440
CV52D003	28	0	0	0	2056	2056
CV52D005	114	0	0	0	1653	1653
CV52D030	23	0	0	0	4743	4743
CV52D019	17	0	0	0	5267	5267
CV52D018	19	0	0	0	5183	5183

VELOCITY SOFTWARE

Using MFC to get more capacity

Start at ESAMFC… (SMT disabled, medium cpus ½ effective)

SET SRM UNPARKING LARGE | MEDIUM

Hyperdispatch medium (**combine 3:1 gives back ONE IFL Capacity**)

Report: ESAMFC MainFrame Cache Magnitudes

		<CPU Busy>		<-------Processor------>				buffer(TLB)-	
		<percent>		Speed/<-Rate/Sec->				CPU Cycles	
Time	CPU	Totl	User	Hertz	Cycles	Instr	Ratio	**Cost**	Lost
15:03:02	0	97.3	89.1	5000M	4856M	1565M	3.103	60.09	2918M
	1	97.3	90.8	5000M	4860M	1469M	3.309	63.35	3079M
	2	97.2	90.5	5000M	4856M	1512M	3.211	63.46	3082M
	3	96.9	89.5	5000M	4842M	1388M	3.487	63.92	3095M
	4	97.1	90.3	5000M	4847M	1655M	2.928	58.78	2849M
	5	**66.0**	**59.8**	5000M	3286M	575M	**5.718**	68.46	2250M
	6	**65.9**	**59.9**	5000M	3283M	597M	5.503	67.38	2212M
	7	**66.0**	**60.1**	5000M	3288M	559M	5.880	68.93	2266M
	12	65.9	59.9	5000M	3282M	544M	6.030	68.45	2247M
	13	65.8	59.7	5000M	3275M	531M	6.173	69.60	2280M
System:		947	868	5000M	47.2G	11.5G	4.120	65.25	30.8G

VELOCITY SOFTWARE

Where do my cycles go?

For z13, 5,000M cycles / cpu / second

Total used (97%): 4856M

Total Miss cost: 4224M

TLB: 2918M (single DAT) ←

L1 Instruction penalty: 2580M (concurrent)

L1 Data Penalty: 3013M (concurrent)

For workload, 600M cycles/sec used for real work?

- 1500M instructions per second in 600M cycles!!!

(For z14, few data samples, low utilization – non conclusive)

VELOCITY SOFTWARE

Cache Sizes – z13

- **L1: 96K Instruction, 128K Data**
- **L2: 2MB Instruction, 2MB data**
- **L3: 64MB (Chip, Shared over 8 CPUS)**
- **L4: 480MB + 224M NIC (per node) (for 24 cpu)**
- **(15000 dispatches / second / thread)**
- **(120,000 dispatches / second / L3)**

Dispatches per second per CPU: 10,000

- 256 byte cache line
- L3: 64MB, supports 270,000 cache lines
- L3 supports 8 cpus, 120,000 dispatches per second

VELOCITY SOFTWARE

Cache analysis arithmetic

Loads for memory per second for L3

- .146 (loads per 100 instructions) / 100
- 1.5B instructions per second
- * 8 cpus per L3
- = 17M cache loads per second

17M cache loads per second / 270,000 cache lines

- Flushes L3 cache 65 times per second, every 15ms
- Polling is every 10 ms. So system tuned for polling
- Real work would never have L3 cache

VELOCITY SOFTWARE

SMT Throughput Thoughts

One execution unit per IFL, SMT helps? Or not?

- Two threads - Neither will get 100% (70%?)
- “idle time” on execution unit when thread takes cache miss

Objective of Multithread (MT)?

- increase Instructions (Capacity) executed on CPU
- Estimates are 20% (Canadian marketing) to 30% (lab)
  (Canadian marketing more accurate than lab….)

z14 includes a new translation lookaside buffer (TLB2) design with four hardware-implemented translation engines that reduces latency when compared with one pico-coded engine on z13

IF 60% of cycles used for DAT single thread on z13,

- **z14 SHOULD BE AWESOME**

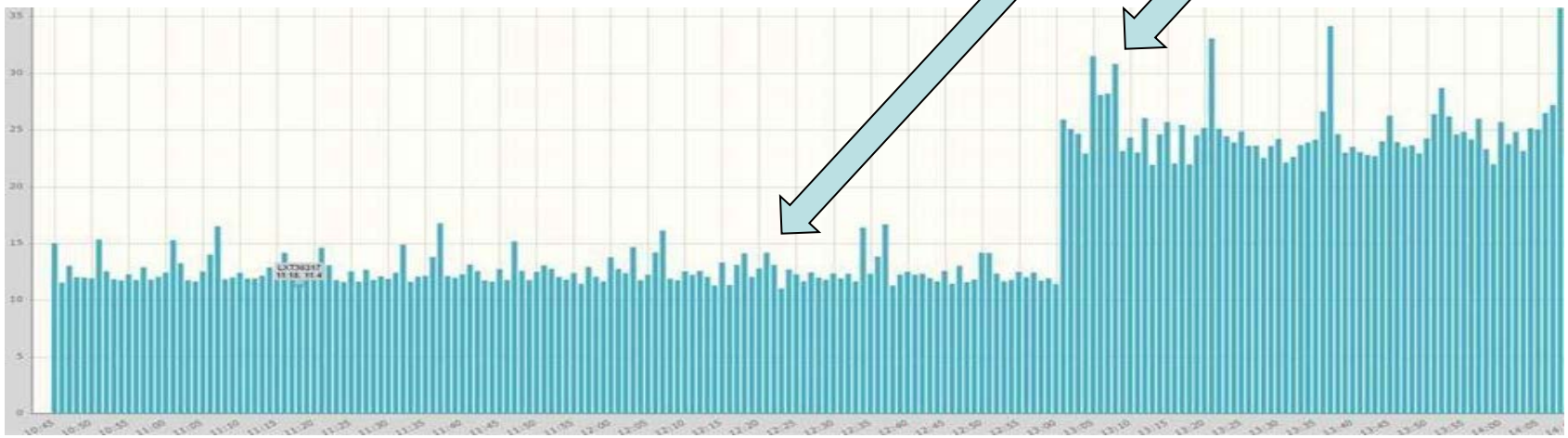
Conclusion: understand MFC before enabling SMT:

VELOCITY SOFTWARE

Customer test – specex off (z13)

This is why you should pay attention!!!!

- **What happens to Linux CPU when SPECEX OFF?**
- **Do you want to understand best options?**
- **What happens when SMT turned off?**
- **How is your capacity plan?**



VELOCITY SOFTWARE

z13 data SMT / SPECEX "test"

Cycles per Instruction – lower is better

- Disabling SMT – "measured cpu utilization goes down"?
- Specex off – instructions double? CPU doubles?
- **7 IFLs in LPAR, For SMT, 14 threads, Low utilization effects**

```
Report: ESAMFC        MainFrame Cache Magnitudes Report
             <CPU Busy> <-------Processor------>
             <percent>  Speed/<-Rate/Sec->
Time      CPU Totl User  Hertz Cycles Instr Ratio
-------- --- ---- ----  ----- ------ ----- -----
System:       170  162  5000M  8151M 2691M 3.029
System:       131  125  5000M  6177M 1651M 3.742
System:       149  141  5000M  7076M 2236M 3.165
System:      86.0 79.2  5000M  4151M 3157M 1.315 ←SMT OFF, 12
System:      91.8 84.0  5000M  4443M 3450M 1.288
System:      81.8 75.3  5000M  3938M 2844M 1.385
System:      66.1 61.9  5000M  3155M 2316M 1.362
System:       111  101  5000M  5403M 4423M 1.222 ←SPECEX OFF
System:       124  112  5000M  6034M 5011M 1.204
System:       102 93.7  5000M  4936M 3973M 1.243
```

VELOCITY SOFTWARE

z14 data SMT / SPECEX "test"

Different Installation, successful SMT user "SAP on dedicated IFLs works great"….

"From yesterday's Velocity ESAUSP5 report. Turning off speculative execution in an SMT-2 environment on a z14 from 10:15 to 11:14 resulted in ~75% increase in CPU use."

VELOCITY SOFTWARE

z13 data SMT / SPECEX "test"

Cycles per Instruction – lower is better

- Disabling SMT – "measured cpu utilization goes down"?
- Specex off – instructions double? CPU doubles?
- **7 IFLs in LPAR, For SMT, 14 threads, Low utilization effects**

```
Report: ESAMFC        MainFrame Cache Magnitudes Report
-------------------------------------------------------------------------
             <CPU Busy>  <-Problem State--> <Level 1 cache/second->
             <percent>   <-Rate/Sec->       Instruction <---Data-->
Time     CPU Totl User   Cycles Instr Ratio Wrtes  Cost Writes Cost
-------- --- ---- ----   ------ ----- ----- ------ ---- ------ ----
System:       131  125    3253M  960M 3.387   2M   372M 7831K  545M
System:       149  141    3696M 1315M 2.810   6M   421M 9552K  618M
System:      86.0 79.2    2390M 2163M 1.105   26M  444M 9932K  742M
System:      91.8 84.0    2581M 2402M 1.075   28M  480M   11M  796M
System:      81.8 75.3    2219M 1878M 1.181   26M  444M 9776K  726M
System:      66.1 61.9    1712M 1435M 1.193   17M  311M 6689K  501M
System:       111  101    3235M 3220M 1.004   36M  564M   14M  993M
System:       124  112    3608M 3659M 0.986   40M  633M   16M 1155M
```

VELOCITY SOFTWARE

What Else is Important?

More instructions per second with static CPU speed

- Pipelineing (concurrent instruction execution)
- Bigger Cache
- Localizing work to cache
- SMT

LPAR – HiperDispatch

- attempts to align Logical CPs with PUs in same Book

Vertical vs Horizontal Scheduling

Affinity

Most of this breaks in a Linux environement

VELOCITY SOFTWARE

Nesting Steals – Affinity (NOT) working?

EC12, 80 IFLs

LPAR: 32 IFLs (p210)

Report: ESAPLDV Processor Local Dispatch Vector Activity

Time	CPU	<VMDBK Steals	Moves/sec> To Master	Dispatcher Long Paths	<-CPU Steals fr <-From Nesting Same	NL1	NL2
14:06:00	0	3529.8	11.6	13104.2	1951	1198	380
	1	2908.6	0	11452.0	1626	976	306
	2	2751.9	0	10475.2	1630	855	267
	…						
	8	3156.8	0	11949.7	1462	1366	329
	9	2702.0	0	10806.9	1283	1137	282
	10	2504.7	0	9849.8	1287	970	248

Steals: vmdblks moved to processor

Dispatcher Long paths:

- **vmdblks dispatched (10K/Sec/CPU)**

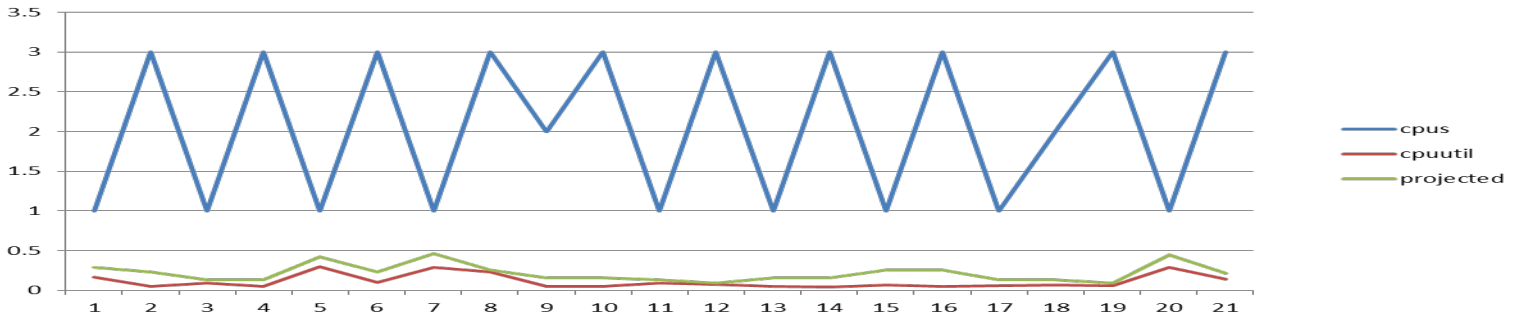
Nesting level – CPU on chip,

- **different chip(NL1),**
- **different book(NL2) (zero on z13…)**

VELOCITY SOFTWARE

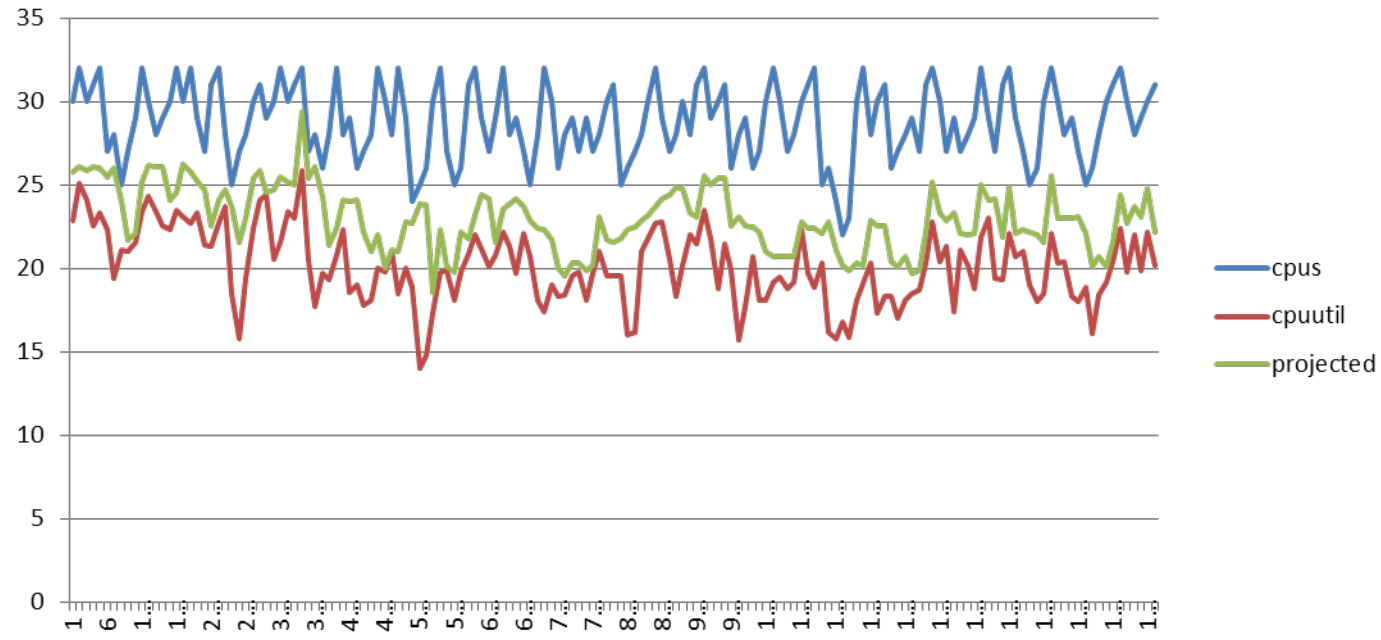
“Apparently, I have a condition, it is too many "vertical-low logical cores". I wish I could take a pill and make them go away...sigh...”

Objectives of “vertical” scheduling
Localize work to cache
Monitor “event” – see parked cpus every 2 seconds
Small number of engines results in high CPU queue



VELOCITY SOFTWARE

Objectives of “vertical”
Localize work to cache
Can we validate this has value?



VELOCITY SOFTWARE

Conclusions – z14

Cycles per Instruction matters

- Reducing cycles per instruction improves capacity
- EFFECTIVE Use of cache (all levels) has positive impact

DAT not a bottleneck on z14

- 20-40% more cycles available for real work
- SMT will have more value
- Need production z14 data at high utilization to validate…
- L1/L2 cache still thrashed, but more effective .

Would be nice if IBM fixed….

- websphere / db2 polling!
- Dispatching 50,000 times per second per thread ridiculous

VELOCITY SOFTWARE

Conclusions – SMT on z13

Cycles per Instruction matters

- Reducing cycles per instruction improves performance
- Use of cache (all levels) has positive impact

SMT on z13?

- Doubles demand for cache on z13,
- TLB Single threaded on z13
- Result: CPI increases so much, capacity drops….

SMT on z13 Effective when:

- Lower dispatch rates (SAP, Oracle)
- Dedicated IFLs to avoid impact from “other workloads”

VELOCITY SOFTWARE

Conclusions – Vertical vs Horizontal

Vertical Objectives:

- Be nice to others (other LPARs)
- Localize cache

Difficult to show value

- Workloads with high dispatch rates don’t reuse cache
- At low number of engines (HiperDispatch), cpu queue goes up, cache hit goes down, value?

VELOCITY SOFTWARE

Performance Questions we have

FPC/EDEV

- What is their performance?
- Show which is better configuration with NUMBERS!!
- NEW Monitor Records in 6.4+

GPFS

- Alert on file space does not work with this file system
- GPFS has it's own mib (ESAGPFx)

Cloud / disks

- How big are these disks that my storage people gave me?
- New metric shows "max blocks/cylinders" (ESADSD1)

VELOCITY SOFTWARE

User diagnose rate from CP Monitor: ESAUSRD / ESALNXG

```
Report: ESAUSRD      Virtual Machine Vis                ZM
-------------------------------------------------------------
UserID  Total <------------diag counts / second-------
/ClassID rate  000  004  008  00C  010  044   09C  0A0
-------- ---- ---- ---- ---- ---- ---- ----  ---- ----
11:04:00  981  0.4 23.4  1.9  0.4  0.1  754   160    .
TCPIP     0.4    .    .    .    .    .    .     .    .
 ***User Class Analysis***
ZVPS     33.7  0.1    .  0.4    .    .    .     .    .
TheUsers  947  0.4 23.4  1.6  0.4  0.1  754   160    .
 ***Accounting Code Analysis**
LXT46410  132    .    .    .    .    .  0.2   132    .
LXT46420  1.6    .    .    .    .    .    .   3.4    .
LXT47401  364    .    .    .    .    .  364   1.8    .
ZALERT    5.6  0.1    .  0.8  0.0    .    .     .    .
 ***CPU POOL User Analysis***
INSTALL   7.1    .    .    .    .    .    .  15.4    .
MQFMPOOL  8.5    .    .    .    .    .  0.0   8.5    .
XDRPOOL   753    .    .    .    .    .  753   8.5    .
 ***Top User Analysis***
LXT46410  132    .    .    .    .    .  0.2   132    .
LXT47401  364    .    .    .    .    .  364   1.8    .
LXT46403  8.3    .    .    .    .    .  0.1   8.3    .
LXT46407  0.5    .    .    .    .    .    .   8.8    .
```

```
; Note: in decimal,

rptdiag(01) = 00 ; 000
rptdiag(02) = 04 ; 004
rptdiag(03) = 08 ; 008
rptdiag(04) = 12 ; 00C
rptdiag(05) = 16 ; 010
rptdiag(06) = 20 ; 014
rptdiag(07) = 36 ; 024
rptdiag(08) = 68 ; 044
rptdiag(09) = 92 ; 05C
rptdiag(10) = 96 ; 060
rptdiag(11) = 100; 064
rptdiag(12) = 104; 068
rptdiag(13) = 124; 07C
rptdiag(14) = 136; 088
rptdiag(15) = 152; 098
```

VELOCITY SOFTWARE

FCP – What do we know? (new IODFCS record)
New reports: ESAEDEV, ESAFCP

```
Report: ESAFCP        FCP Emulated Device (EDEV) Report                Software Cor
-------------------------------------------------------------------------------
Date/     <-FCP Device-> Path <IO/Second-> <MB/Second>  <QueueDepth> QDIO QTime
Time      No.    Type    Cnt  Reads Writes Reads Writes   Avg  Avg   Microsecs
-------- ----  -------- ---- ----- ------ ----- ------ ----- -----  ---------
10:40:00 5100  ficonE8S    4   1.5    4.9   0.0    0.0   1.0   1.0     2983.4
         5700  ficonE8S    5   0.2    4.8     0    0.1   1.0   1.0    10954.8
```

```
Report: ESADEV1       Device Configuration (non-DASD)
-------------------------------------------------------
 Dev       Device <CHPIDs OnLn> OBR <-Cntrl Unit-> UserID
 No. SysID Type   01 02 03 04   Code Code Model    (if ded)
---- ----- ------ -- -- -- --   ---- ---- -------  --------
5100 0033  1732-3 21 .  .  .     00   00  1731-3    .
5101 0034  1732-3 21 .  .  .     00   00  1731-3    .
5102 0035  1732-3 21 .  .  .     00   00  1731-3    .
```

VELOCITY SOFTWARE

EDEV – What do we know? (new IODCHS record)

```
Report: ESADSD1      DASD Configuration
-----------------------------------------------
 Dev Sys          Device    <CHPIDS OnLn><-Cntrl Uni
 No. ID   Serial  Type   SHR 01 02 03 04 OBR/CU Mode
---- ---- ------ ------ --- -- -- -- -- ------ ----
1001 0033 640RL1 9336   NO  21  .  .  .  00/00 6310
1002 0033 M01RES 9336   NO  21  .  .  .  00/00 6310
1003 0033 M01S01 9336   NO  21  .  .  .  00/00 6310
1004 0033 M01P01 9336   NO  21  .  .  .  00/00 6310

Report: ESAEDEV      EDEV Emulated Device (EDEV) Report             Veloc
-------------------------------------------------------------------------
Date/    <EDEV Channel Path> <IO Ops/Sec> <MegaB /Sec> <Pct Utilization>
Time     ID  Type       count Reads Writes Recivd Sent   CPU   Bus Adapte
-------- -- --------- ----- ----- ------ ------ ----  ---- ---- -------
10:40:00 20  Fabric        1   0.2    0.9      0  0.0     0    0       0
         21  Fabric        4   1.5    5.0    0.0  0.0     0    0       0
         27  Fabric        5   0.2    4.8      0  0.1     0    0       0
```

VELOCITY SOFTWARE

GPFS: Data from snmp – problem? How full….

```
Report: ESAGPFS      GPFS Cluster File System Config                Velocity
-------------------------------------------------------------------------------
Collector                                                  Node FS
Node     Cluster Name              GPFS ID           Rlse Cnt  Cnt   Domain
-------- ------------------  -------------------- ---- ---- ----  --------
11:56:00
ssnode1  cluster1.ssnode1     5049816574407790568  1700    3     1  cluster1

Report: ESAGPFSN     GPFS File system Configuration                 Velocity
-------------------------------------------------------------------------------
Collector                                Plat-                        Thread
Node     Idx  Name    IP Address        Form Status Fails  Wait  Good Versn
-------- --- -------- --------------- ---- ------ -----  ----  ----  ----
11:56:00
ssnode1   49 ssnode1  192.168.5.92    S390 up      0      yes  none 4.2.3.6
          50 ssnode2  192.168.5.93    S390 up      0      yes  none 4.2.3.6
          51 ssnode3  192.168.5.94    S390 up      0      yes  none 4.2.3.6
```

VELOCITY SOFTWARE

GPFS: Data from snmp

```
Report: ESAGPFSS     GPFS Storage Pool Configuration
-------------------------------------------------------
Collector Subpool  Files
Node      Name     System   Storage   Free Disks
--------  -------- -------- ------- ------ -----
11:56:00
ssnode1   system@@ gpfs1@@@    192K   185K     0


Report: ESAGPFSD     GPFS DISK Configuration/Analysis
Monitor initialized: 06/22/18 at 11:54:12 on 2828 serial 0314C7
-------------------------------------------------------------------
Collector                  StgPool Disk  <Dsk Blks> Sub  <I/O Time>
Node      DiskName FSName  Name    Status Total Free free Read Write
--------- -------- -------- ------ ------ ----- ---- ---- ---- -----
11:56:00
ssnode1   disk1    gpfs1    stem   InUse 192352 185K 13.7 1.1M     0
```

VELOCITY SOFTWARE

Futures for “Next” Share

Docker / Kubernetes
IPV6 Full support
z/OS???
DB2
MQ
ILMT

VELOCITY SOFTWARE

Thank You

Please Send data for z14, FCP/SCSI, GPFS

Please send performance problems (raw monitor data, zvps history data)

SPECIAL WEBSITES….

- **VelocitySoftware.com/HANDOUTS**
- **VMWORKSHOP.ORG (140 Real Attendees… June 26-29)**
- **velocitysoftware.com/seminar/workshop.html**

VELOCITY SOFTWARE