Performance Management for 1000 virtual servers

(SHARE SESSION 10335)

Barton Robinson
Velocity Software
Barton@VelocitySoftware.com

VELOCITY SOFTWARE

PROVEN PERFORMANCE

Performance Management for 1000 virtual servers

Objectives

- Define service management
- How to define service targets
- Drill down requirements
- VM Subsystems
- Linux Subsystems
- Alerts Discussions

- OR, avoid problems, spend an extra “x” million “cash units” to avoid problems

Velocity Software Performance Management

- **Instrumentation Requirements**
  - **Performance Analysis**
  - **Operational Alerts**
  - Capacity Planning
  - Accounting/Charge back
- **Correct data**
  - (rmfpms is being withdrawn, SOD)
- **Capture ratios**
- **Instrumentation can NOT be the performance problem**

Performance Analysis Skills Requirement

Performance skills pay for themselves

- Buy 16 gb of storage, or “set mdc”?
- Buy more IFLs or more DASD Cache?
- Buy more CECs or reduce storage requirements
- What solution does a hardware vendor offer, tuning or….

Free performance class? (Mountain View, CA)

ZTUNE anyone? (performance service offering)

Performance Analysis for 1000 virtual servers

Today's requirements:

- 1,000 virtual servers under z/VM?
- 10,000 virtual servers under ESX
- 5,000 P-series

Performance analysis requirements

- Hypervisor layer (LPAR, ESX)
- Server layer
- Application layer
  - (cache one call, reduce overall response time by 50%)
- **Full data capture (crystal balls not needed)**
- Useable granularity

Full data capture to diagnose problems

- To resolve, requires:
  - Linux Process data before and after across ALL servers
  - Virtual machine data before and after
  - Disk performance data before and after

```
Report: ESALPARS     Logical Partition Summary
Monitor initialized: 02/06/07 at 13:00:00 on 2094 serial 2BFBD
-------------------------------------------------------------------
         <--Complex--> <-------Logical Partition--->  <-Assigned
          Phys Dispatch                Virt <%Assigned>  <---LPAR-->
Time      CPUs    Slice Name      Nbr CPUs Total  Ovhd  Weight  Pct
--------  ---- -------- -------- --- ---- -----  ----  ------ ----
14:17:00    13  Dynamic Totals:    0   30 791.3  11.1     100  100
14:18:00    13  Dynamic Totals:    0   30 782.6   9.6     100  100
14:19:00    13  Dynamic Totals:    0   30 857.4  11.4     100  100
14:20:00    13  Dynamic Totals:    0   30 866.0  10.7     100  100
14:21:00    13  Dynamic Totals:    0   30 866.3  10.8     100  100
14:22:00    13  Dynamic Totals:    0   30 811.2  10.9     100  100
14:23:00    13  Dynamic Totals:    0   30 809.6  10.7     100  100
14:24:00    13  Dynamic Totals:    0   30 876.3  11.4     100  100
14:25:00    13  Dynamic Totals:    0   30 834.5  10.7     100  100
14:26:00    13  Dynamic Totals:    0   30  1187   8.8     100  100
14:27:00    13  Dynamic Totals:    0   30  1286   5.5     100  100
14:28:00    13  Dynamic Totals:    0   30  1282   5.5     100  100
14:29:00    13  Dynamic Totals:    0   30  1293   3.5     100  100
```

Performance Analysis for 1000 virtual servers

Reality check

- 100-150 servers per z/VM LPAR
- ESX, MS, P unlimited

- System level problems solved at a system level

- "VM SYSPLEX" Announced:
  - Different set of challenges

Performance Analysis for 1000 virtual servers

Analyzing 1,000 servers

- Concepts are not new
- Traditional sites ran 5,000 concurrent logged users (virtual machines)
- Requirements for analysis are
  - "the same but different"

Operational alerts required to alert

From z/VM, from Linux

Performance analysis high level granularity,

- Zoom to finer granularity

Performance Management implies Service Contracts, Chargeback, and SLAs

Service contracts based on

- Keeping users happy
- Financial obligations

Requirements

- Detect service issues
- Resolve service issues - QUICKLY

Traditional CMS (one process per virtual server)

- CMS Response times
- Percentiles

Linux Response time

- Applications require instrumentation (ARM)
- Cost of ARM currently prohibitive (opportunity)

Linux Load measurements more common

- By application, server

Detecting service issues requires

- Defining your workload in manageable units
- Defined normal operational parameters
- Record history of application, servers

Define “application” or “server” norms by

- CPU trends
- Storage requirements
- I/O requirements
- Network activity (bandwidth, connections)

Performance Analysis Platform Requirements

Platforms have specific requirements

z/VM Subsystems

- LPAR Analysis, Processor
- Storage / Paging
- DASD
- Scheduler / Dispatcher

Linux Servers

- Storage / Swap
- Process Table

Other virtualization platforms have different requirements

Performance Analysis Large System Challenges

DASD Subsystem Analysis

- Analyze performance problems for 5,000 disks?
- Some installations have 20,000+ (20 years ago)
- Now we have “Shared DASD” that needs to be measured

Virtual Machines

- Analyze performance for 1,000 virtual machines
- 5,000 was common 20 years ago
- (150 per lpar is a reasonable maximum today)

Linux Servers

- Analyze performance for 1,000 nodes
- How many on one “LPAR” or “ESX”?
- How many LPARs or ESX Servers?

Large system? Concepts?

1,119 users logged on, 180 active

```
Screen: ESAMAIN  Velocity Software - VSIVM4      ESAMON 3.
1 of 3  System Overview                          LIMIT 500

         <---Users----> Transact.      <Processor>  Cap-
         <-avg number->  per Avg.      Utilization  ture
Time       On Actv In Q Sec. Time CPUs Total Virt. Ratio
*------- ---- ---- ---- ---- ---- ---- *---- ----- -----
10:01:00 1119  183 13.0 29.6 0.13    1  31.4  28.0   100
10:00:00 1119  179 19.0 30.8 0.11    1  15.0  13.8   100
09:59:00 1119  181 18.0 30.1 0.10    1  15.0  13.7   100
09:58:00 1120  179 15.0 29.9 0.11    1  14.9  13.6   100
09:57:00 1120  149 19.0 29.3 0.11    1  15.3  14.0   100
09:56:00 1119  120 19.0 28.8 0.13    1  61.0  59.7   100
09:55:00 1119  123 20.0 28.3 0.11    1  14.9  13.7   100
```

VM “granularity”

1000 servers require grouping

Traditional “O/V”

- User (virtual machine) classification
  - by function,
  - users by department

Linux servers (Grouping requires working knowledge)

- Group by application (http, printing, database)
- Department
- Application subsystem

Virtual Machine "granularity" Requirements

Group virtual machines on one LPAR

Group "nodes"

- By application
- By ESX

Requirements to group applications across nodes

- Understand loads when workload balanced across many servers

Requirements to monitor ESX, BX, Blades

Analyze 1000 servers with classification “USERCLASS”

```
Screen: ESAUSP2  Velocity Software     ESAMON 3.808 08/05 09:45-09:46
1 of 3  User Percent Utilization       CLASS *               2096 44B42
                                <-------Main Storage-------->
          UserID    <Processor> <Resident->  Lock <-WSSize-->
Time      /Class    Total  Virt Total  Actv   -ed Total  Actv
--------  --------  ----- ----- ----- ----- ----- ----- -----
09:46:00  System:   65.82 64.90  604K  558K  6407  611K  559K
          SUSE      48.22 48.11  150K  150K  1475  148K  148K
          TEST       5.70  5.52  190K  180K  2102  188K  178K
          *TheUsrs   4.81  4.76 23636 23627   528 24671 23084
          REDHAT     3.04  3.00  176K  176K   877  182K  182K
          Velocity   2.99  2.90  7235  6861    21  7855  6834
          KeyUser    0.66  0.29  3486  3486  1332  2239  2154
          Servers    0.26  0.20  1786  1026     6  1947  1022
          ORACLE     0.12  0.12 15962 15962    66 15875 15875
          DEMOTH     0.01  0.00 37409  2255     0 41168  2255
```

Zoom on "CLASS SUSE" shows what is important

```
Screen: ESAUSP2  Velocity Software      ESAMON 3.808 08/05 09:45-09:46
1 of 3  User Percent Utilization        CLASS SUSE USER *    2096 44B42

                                <-------Main Storage-------->
          UserID    <Processor> <Resident->  Lock <-WSSize-->
Time      /Class    Total  Virt Total  Actv   -ed Total  Actv
--------  --------  ----- ----- ----- ----- ----- ----- -----
09:46:00  SUSELNX2  45.51 45.47  4198  4198     0  4080  4080
          SLES11X    0.20  0.20 11809 11809   214 11595 11595
          SLES9X     0.17  0.16 12182 12182    35 12147 12147
          SLES9      0.15  0.15 13012 13012   208 12804 12804
          SLES8X     0.10  0.09 35221 35221   250 34950 34950
          SLES8      0.07  0.04 28332 28332     0 33975 33975
          SLES10     0.05  0.05 23646 23646   219 23427 23427
          SLES10X4   0.04  0.04 44438 44438   229 44209 44209
          SUSEAPPS   0.02  0.02 11503 11503    11 11492 11492
```

User classification is worth the investment

DASD Subsystems – analysis for 4,000 3390s?

- Disk configuration shows 4,000 devices, granularity issue

```
 ESADSD1  LISTING  A1  V 129  Trunc=129 Size=9344 Line=0 Col=1 Alt=0
====>
00000 * * * Top of File * * *
00001 1Report: ESADSD1      DASD Configuration
00002  Monitor initialized: 11/06/10 at 16:07:10 on 2097 serial 374E2
00004  ------------------------------------------------------------
00005   Dev Sys          Device    <CHPIDS OnLn><-Cntrl Unit-> UserID
00006   No. ID   Serial  Type   SHR 01 02 03 04 OBR/CU Model   (if ded
00007  ---- ---- ------ ------ --- -- -- -- -- ------ ------- -------
00008
00009  0400 0000 LT35C2 3390-9 YES B4 D9 BA BB  3C/00 2107
00010                           D4 D5 E6 DB
00011  0401 0001 LT35C3 3390-9 YES B4 D9 BA BB  3C/00 2107
00012                           D4 D5 E6 DB
00013  0402 0002 LT35C4 3390-9 YES B4 D9 BA BB  3C/00 2107
00014                           D4 D5 E6 DB
00015  0403 0003 LT35F6 3390-9 YES B4 D9 BA BB  3C/00 2107
00016                           D4 D5 E6 DB
00017  0404 0004 LT35F7 3390-9 YES B4 D9 BA BB  3C/00 2107
00018                           D4 D5 E6 DB
```

DASD Subsystems – analysis for 4,000 devices?

- What is relevant for performance? (top 10 busy devices)

```
Report: ESADSD2      DASD Performance Analysis                        Linux Te
-------------------------------------------------------------------------------

                                                        <--------DASD Response
 Dev         Device <--SSCH--> <%DevBusy> <SSCH/sec->          <--Service time
 No. Serial Type    Total  ERP  Avg  Peak   avg  peak  Resp   Serv Pend Disc
---- ------ ------ ------ ---  ---- ---- ----- ----- -----   ---- ---- ----
16:09:07
***Top DASD by Device busy***
BC15 LP394B 3390-9  40334   0  44.0 44.0 672.2 672.2   0.7    0.7  0.3  0.0
EC6D LP366D 3390-9  12495   0  36.6 36.6 208.3 208.3   1.8    1.8  0.2  0.1
EF05 LP3505 3390-9   2741   0  26.3 26.3  45.7  45.7   5.7    5.7  0.3  3.2
B908 LP3B84 3390-9  12988   0  26.2 26.2 216.5 216.5   1.2    1.2  0.3  0.5
EC4A LP3641 3390-9   8134   0  26.0 26.0 135.6 135.6   1.9    1.9  0.2  1.1
EC2A LP356C 3390-9  18559   0  22.6 22.6 309.3 309.3   0.7    0.7  0.2  0.0
EC29 LP356B 3390-9  15997   0  19.2 19.2 266.6 266.6   0.7    0.7  0.2  0.0
B905 LP3B81 3390-9   9595   0  18.6 18.6 159.9 159.9   1.2    1.2  0.2  0.4
ED12 LP3547 3390-9   1643   0  15.1 15.1  27.4  27.4   5.5    5.5  0.3  2.0
EC49 LP3640 3390-9   4067   0  14.0 14.0  67.8  67.8   2.1    2.1  0.2  1.2
***End Top DASD by Device busy***
```

DASD Subsystems – analysis for 4,000 devices?

- Show control units (256 devices / ctl unit?)

Report: ESADSD2 DASD Performance Analysis Linux Te

									<--------DASD Response			
Dev		Device	<--SSCH-->		<%DevBusy>		<SSCH/sec->			<--Service time		
No.	Serial	Type	Total	ERP	Avg	Peak	avg	peak	Resp	Serv	Pend	Disc
16:09:07												
B800		Control Unit	1881	0	0.0	0.0	31.3	31.3	2.2	2.2	0.2	1.0
B900		Control Unit	28808	0	0.3	0.3	480.1	480.1	1.3	1.3	0.3	0.5
BC00		Control Unit	42348	0	0.3	0.3	705.8	705.8	0.9	0.9	0.3	0.2
BD00		Control Unit	3350	0	0.1	0.1	55.8	55.3	4.3	4.3	0.5	1.7
BF00		Control Unit	1451	0	0.0	0.0	24.2	24.2	4.7	4.1	0.3	2.7
EC00		Control Unit	87867	0	0.9	0.9	1465	1456	1.3	1.2	0.1	0.4
ED00		Control Unit	20413	0	0.9	0.9	340.3	336.6	5.2	5.2	0.7	1.4
EE00		Control Unit	5506	0	0.2	0.2	91.8	91.0	4.9	4.9	0.6	0.9
EF00		Control Unit	4975	0	0.3	0.3	82.9	82.9	4.3	4.1	0.3	2.3
System:			203294	0	0.1	0.1	3389	3343	1.8	1.8	690	62K

Storage Subsystem – analysis for 40 Million pages

```
Report: ESASTR1                                              4.1.0 09/08/11  Page   1
Monitor initialized: 102097                 First record analyzed: 10/09/09 09:22:00
-------------------------------------------------------------------------------
         Users <-------Pages-------------------------------------------------> Over
         Loggd System   <Available> Systm  User  NSS/DCSS  <-AddSpace> VDISK <MDC> Diag Commit Capt-
Time        On Storage  <2gb  >2gb  ExSpc Resdnt Resident  Systm User  Rsdnt Rsdnt  98  Ratio  Ratio
-------- ----- -------  ----- -----  ----- ------ --------  ----- -----  ----- ----- ---- -----  -----
09:23:00   211  40108K    284  5822 11492 38963K     6169   713K     0  2758    12 2769 2.310  0.998
09:24:01   211  40108K    212  3866 11493 38966K     5369   713K     0  2636    12 2769 2.310  0.998
09:25:00   211  40108K    131  3694 11494 38968K     4608   713K     0  1527    12 2769 2.310  0.998
```

Where to analyze?

- Users consuming 39M out of 40M pages

Storage Subsystem – analysis for 39 Million pages

- Which class of "Z/VM" virtual machines?

```
Report: ESAUSPG       User Storage Analysis
-----------------------------------------
         <---Storage occupancy in pages--->
UserID   <---Main Storage---> <--Paging--->
/Class   Total   >2gb   <2GB  Xstor   DASD
-------- ------ ------ ------ ------ ------
09:27:00 38966K 578615 38387K  2593K 64888K
 ***User Class Analysis***
Servers      2      2      0    858   7501
ZVPS      2950   2946      4   2031   5367
Linux      100     96      3     86    389
SOUTHLK  11685K      0 11685K 442949 14670K
APPS     140269      0 140269  16060 342250
NORTHLK   5841K      0  5841K 253513  7802K
WESTCST   3410K      0  3410K  88959  4962K
TheUsers 17883K 571131 17312K  1787K 37057K
```

Nodes are classified into “node groups”

- Which “NODE group” is using CPU?
- Which “esx group” is using CPU?

```
Report: ESAUCD4       LINUX UCD System Statistics Report
Monitor initialized: 09/10/09 at 09:21:50 on 2097
-------------------------------------------------------------------------
Node/                                    <---------Rates Per Second------>
Date/     <Processor Pct Util> NICE <-Swaps-> <-Blocks> Switch Intrpt
Time      Total Syst User Idle Time    In  Out    In  Out    Rate   Rate
-------- ----- ---- ---- ---- ---- ---- ---- ---- ---- ------ ------
09:27:00
***Node Groups***
SOUTHLK    36.0 25.0 11.0 2723    0    0    0    0    0 9644.8 3323.6
APPS        1.0  1.0    0 98.0    0    0    0    0    0  180.0   64.9
NORTHLK    16.0 15.0  1.0 1955    0    0    0    0    0 5539.7 1599.3
WESTCST     7.0  7.0    0 1076    0    0    0    0    0 2753.0  776.9
TheUsers 104.0 91.0 13.0 9768    0    0    0    0    0  20171  14855
```

Node Process "granularity" Requirements

```
Report: ESAHSTA       LINUX HOST Application Report
Monitor initialized: 09/10/09 at 09:21:50 on 2097 serial
-------------------------------------------------------------------------
Node/     Process/    <Application Status Counts> <-----Processor---->
Date      Application             Run-  Res  Load  <---Utilization---->
Time      name        Total Actv ning  Wait  -ed  Percent seconds Avg
--------  ----------- ----- ---- ----  ---- ----  ------- ------- ----
09:23:00
***Node Groups***
SOUTHLK   *Totals*     3136  270 28.0  3103    0    83.8    50.5  0.0
          java       919.0  137    0   914    0    61.3    36.9  0.1
          khelper      5.0  5.0    0   5.0    0     0.6     0.3  0.1
          kjournal   142.0 73.0    0   142    0    12.9     7.8  0.1
          snmpd       28.0 28.0 28.0     0    0     3.3     2.0  0.1
APPS      *Totals*    58.0  8.0  1.0  57.0    0     1.2     0.7  0.0
          java         3.0  3.0    0   3.0    0     0.7     0.4  0.2
NORTHLK   *Totals*    1411  155 20.0  1391    0    35.2    21.1  0.0
          java        97.0 70.0    0  97.0    0    21.9    13.2  0.2
          kjournal   100.0 52.0    0   100    0     6.6     4.0  0.1
          multipat     5.0  5.0    0   5.0    0     5.2     3.1  1.0
          snmpd       18.0 18.0 18.0     0    0     1.0     0.6  0.1
WESTCST   *Totals*   897.0 67.0 11.0   886    0    19.1    11.4  0.0
          java       202.0 26.0    0   202    0    12.1     7.2  0.1
          kjournal    40.0 23.0    0  40.0    0     3.5     2.1  0.1
          multipat     2.0  2.0    0   2.0    0     1.2     0.7  0.6
          snmpd       11.0 11.0 11.0     0    0     1.4     0.9  0.1
```

A scalable z/VM Performance Monitor

Traditional model (1989)

ESAMON: Real time analysis

- Uses Standard CP Monitor

  Real Time Analysis

ESAMAP: Performance Reporting

Post Processing

Creates Long Term PDB

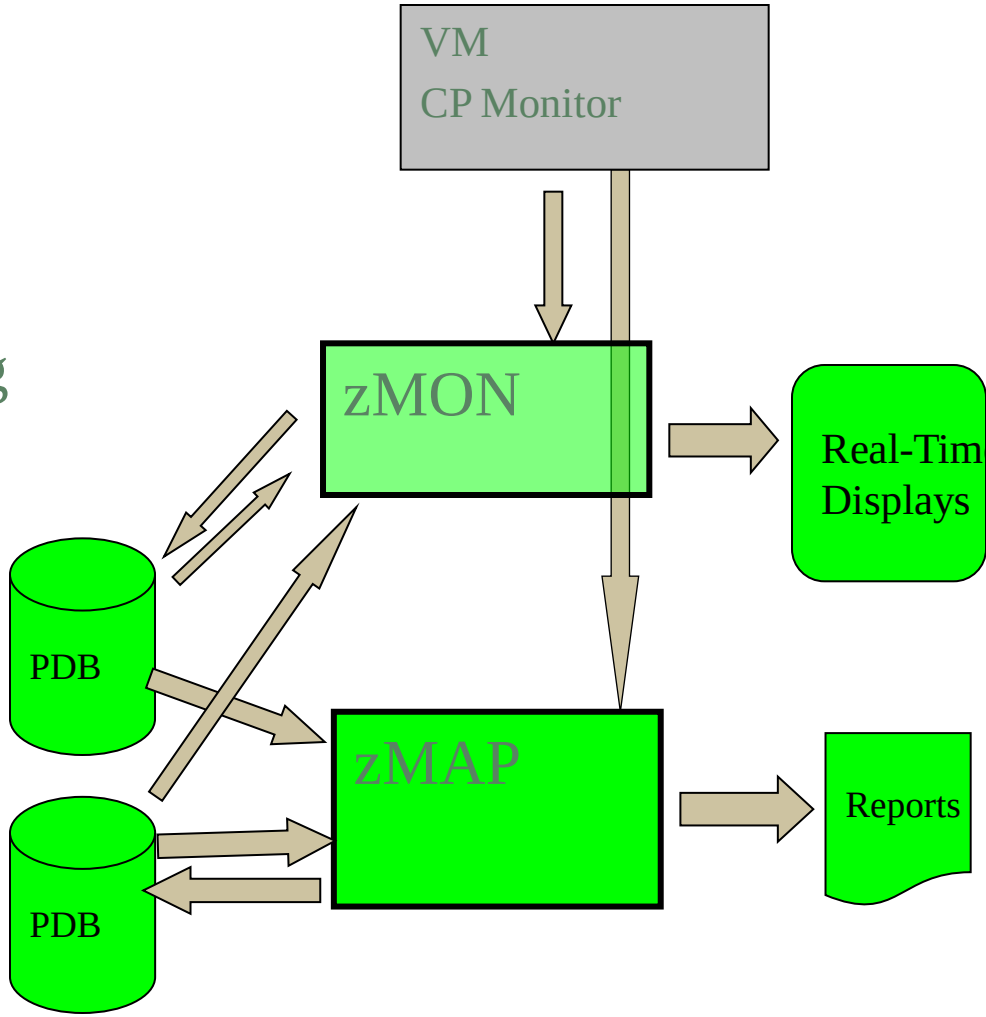
PDB or monwrite data input

PDB (Performance DataBase)

Complete data

By Minute, hour, day

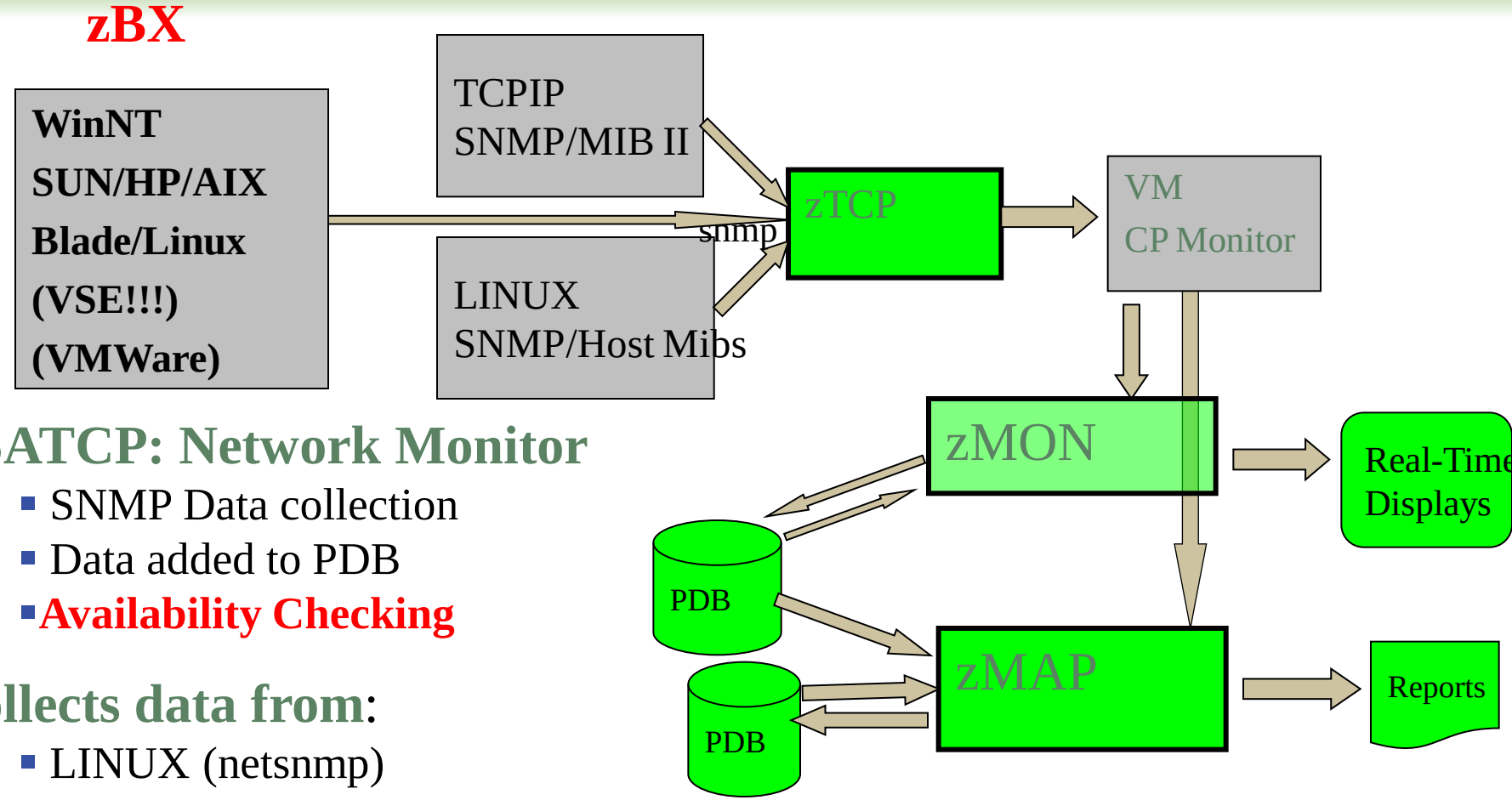
Monthly/Yearly Archive



VELOCITY SOFTWARE

PROVEN PERFORMANCE

Linux and Network Data Acquisition



ESATCP: Network Monitor

- SNMP Data collection
- Data added to PDB
- **Availability Checking**

Collects data from:

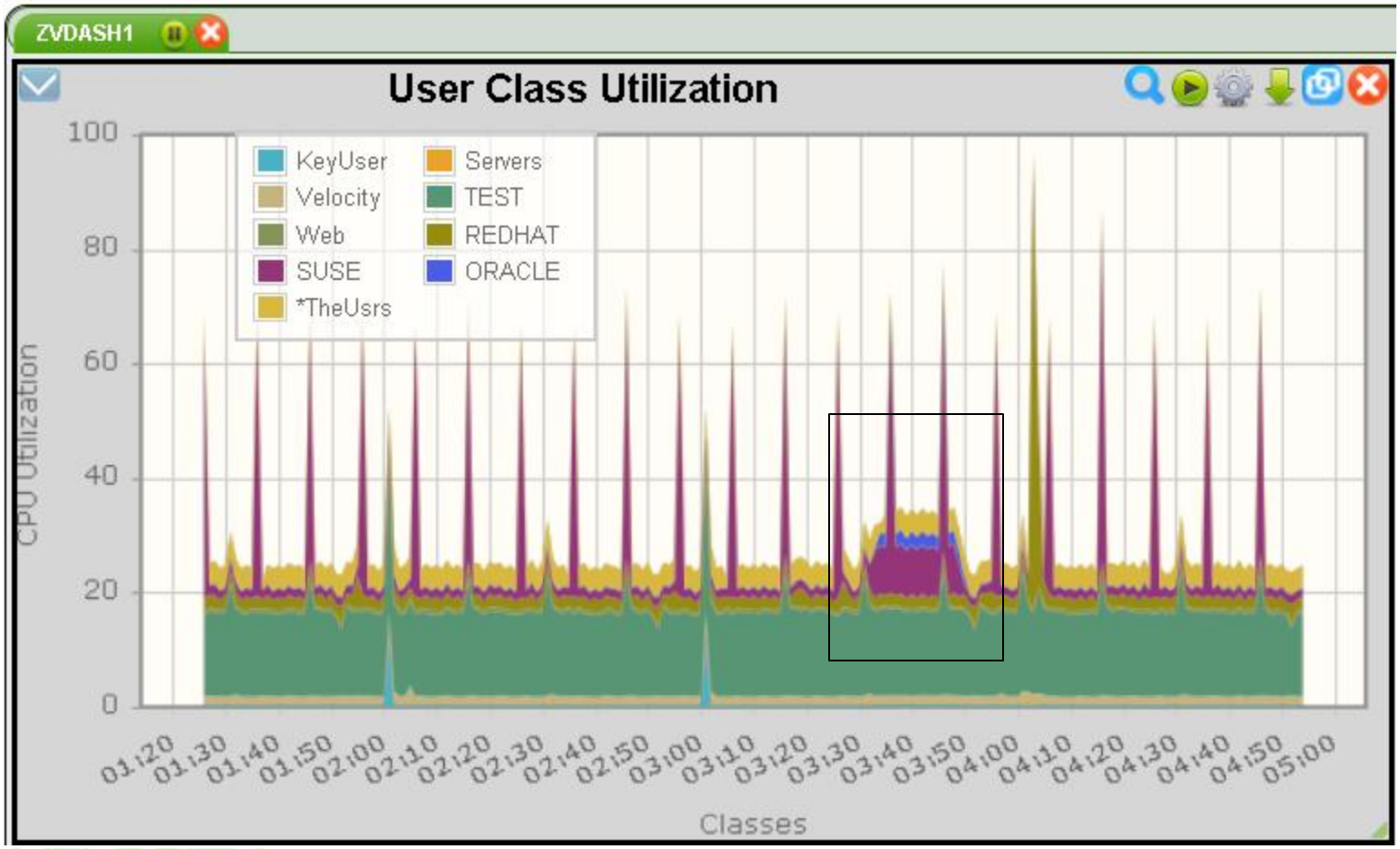
- LINUX (netsnmp)
- NT/SUN/HP (native snmp)
- Printers/Routers….

Using snmpd is VERY scalable!!!

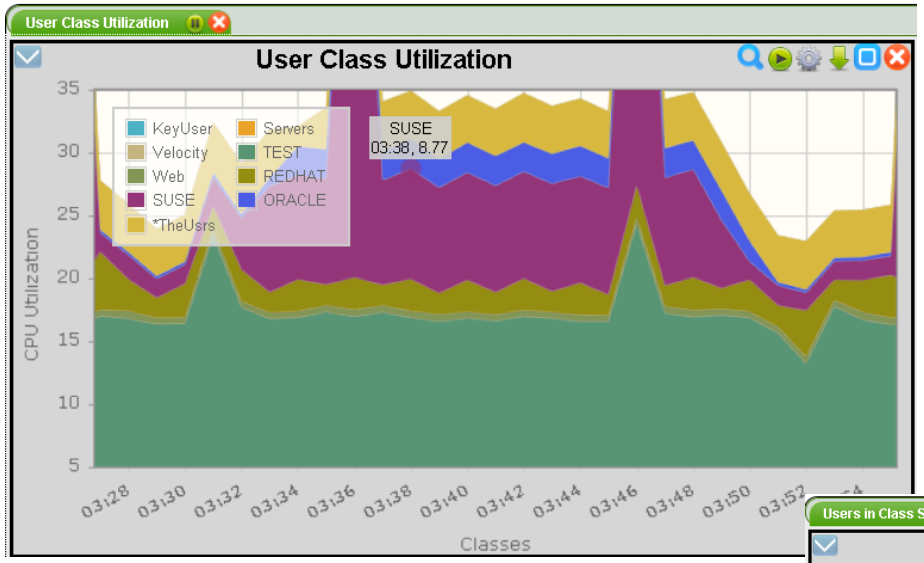
Performance collection for 1000 servers: 30% of one IFL?

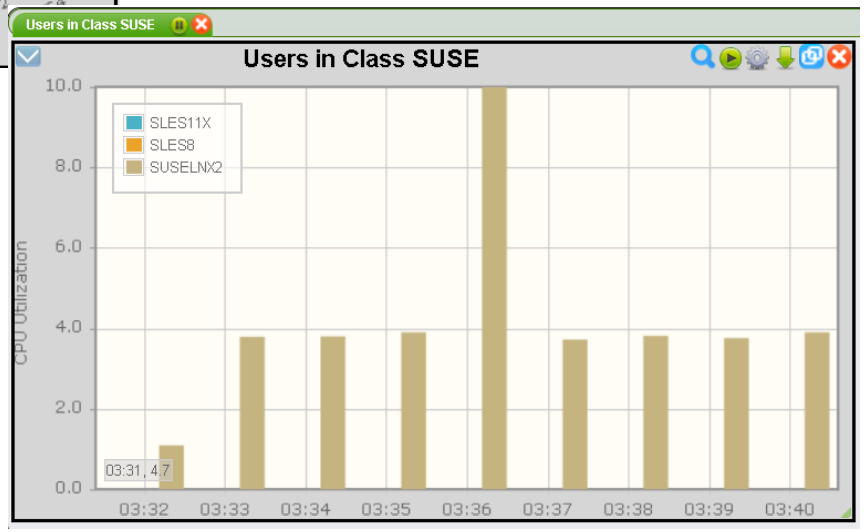
```
Report: ESALNXP      LINUX HOST Process Statistics Report
-------------------------------------------------------
node/     <-Process Ident-> Nice <------CPU Percents---->
 Name      ID     PPID   GRP  Valu  Tot  sys user syst usrt
--------- ----- ----- ----- ---- ---- ---- ---- ---- ----
 snmpd     2290     1  2289  -10 0.02 0.02    0    0    0
 snmpd     1776     1  1775  -10 0.02 0.02    0    0    0
 snmpd     2069     1  2068  -10 0.02    0 0.02    0    0
 snmpd     2193     1  2192  -10 0.25 0.12 0.12    0    0
 snmpd     2166     1  2165  -10 0.04 0.02 0.02    0    0
 snmpd     2230     1  2229  -10 0.04 0.02 0.02    0    0
 snmpd     3370     1  3369  -10 0.03 0.02 0.02    0    0
 snmpd    11202     1 11201  -10 0.04 0.02 0.02    0    0
 snmpd     7855     1  7854  -10 0.02 0.02    0    0    0
 snmpd     5848     1  5847  -10 0.08 0.06 0.02    0    0
 snmpd     2290     1  2289  -10 0.04 0.02 0.02    0    0
 snmpd     1776     1  1775  -10 0.04 0.02 0.02    0    0
 snmpd     2069     1  2068  -10 0.02 0.02    0    0    0
```

Ability to see wide range of data, drill down



Drill down





VELOCITY SOFTWARE

PROVEN PERFORMANCE

1000 Server Disk Space Requirements

One minute performance data

- 4000 cylinders for 1000 servers per day
- 4 cylinders per server per day

Daily Archive

- 300 cylinders for 1,000 servers per day

Disk space “tailorable” with thresholds

VM “granularity” in zVPS

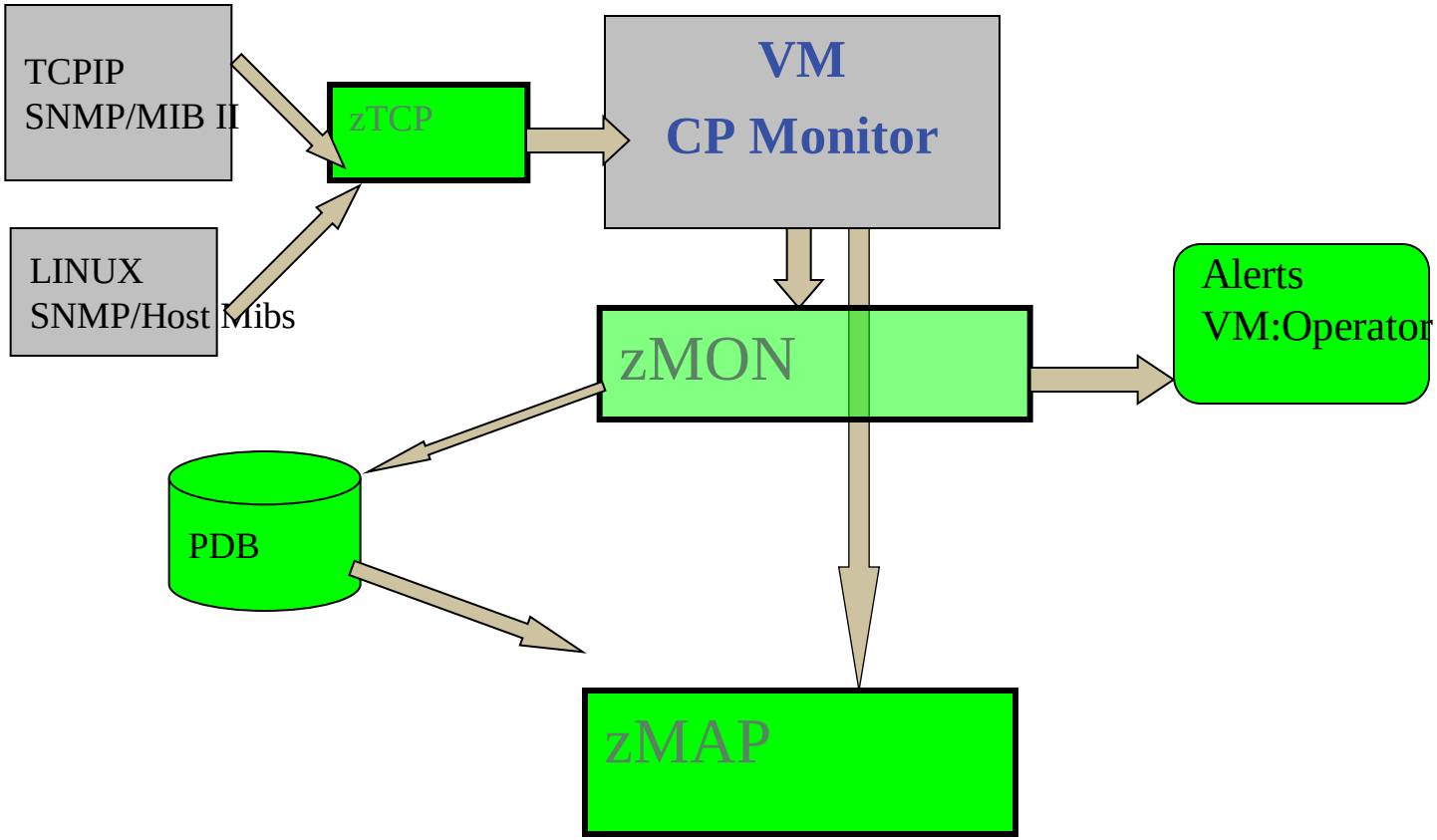
User classification (Virtual Machines)

- USERCLASS(3,x) = “HTTP*”
- USERCLASS(4,x) = “SAP*” ;sap servers
- USERCLASS(5,x) = “Z*” ;Velocity servers

Node classification

- USERCLASS(6,x) = “ESXA*” ; esx server
- USERCLASS(7,x) = “ESXB*”

Operational Alert Support



Performance Analysis Education

Velocity Software's performance workship

- ("free" in April)

VM Workshop – University of Kentucky

SHARE

User groups

Performance Analysis Summary

Analyzing 1000 virtual machines was easy 20 years ago

Analyzing today's Linux servers requires added data

- Linux data
- Other virtual platforms

1000 (10,000) servers requires:

- **Management.**
- **Scalable management tools**
- **No Crystal balls**