A08

AIX Performance Tools update

Luc Smolders

IBM SYSTEM p, AIX 5L and LINUX TECHNICAL UNIVERSITY
Sept 11 - 15, 2006

Las Vegas, NV

Agenda

- Multiple page sizes
  - svmon,vmstat
- NFS V4
  - curt, netpmon
- Tprof enhancements
  - privately loaded and named shared libraries support
  - milicode and hypervisor support
- Hardware PM tools enhancements
  - Power5+ support (5.3 ML4/5.2 ML 8)
  - new set of interfaces reporting time
  - counter multiplexing
  - cpu Dynamic Reconfiguration support
- Automatic performance metric recording
  - local and CEC-wide metrics and reports
  - recording data extraction and links to WLE and nmon_analyzer.
- VIOS Performance monitoring update

Multiple Page Sizes - vmstat

```
# vmstat -P ALL 5

System configuration: mem=2176MB

pgsz            memory                           page
----- -------------------------- ------------------------------------
           siz       avm       fre     re     pi     po     fr      sr     cy
    4K    293136     84593    189715      0      0      0      0       0      0
   64K     16495      1137     15358      0      0      0      0       0      0

    4K    293136     84593    189715      0      0      0      0       0      0
   64K     16495      1137     15358      0      0      0      0       0      0

# vmstat -p ALL 5

System configuration: lcpu=4 mem=2176MB

kthr     memory              page              faults        cpu
----- ----------- ------------------------ ------------ -----------
 r  b    avm   fre    re  pi  po  fr   sr  cy  in   sy  cs us sy id wa
 0  0 102817 435411   0   0   0   0    0   0   4   33 160  0  0 99  0

  psz   avm   fre   re  pi  po  fr   sr  cy     siz
   4K 84593  189715  0   0   0   0    0   0  293136
  64K  1139  15356   0   0   0   0    0   0   16495

kthr     memory              page              faults        cpu
----- ----------- ------------------------ ------------ -----------
 r  b    avm   fre    re  pi  po  fr   sr  cy  in   sy  cs us sy id wa
 0  0 102817 435411   0   0   0   0    0   0   3    8 159  0  0 99  0

  psz   avm   fre   re  pi  po  fr   sr  cy     siz
   4K 84593  189715  0   0   0   0    0   0  293136
  64K  1139  15356   0   0   0   0    0   0   16495

kthr     memory              page              faults        cpu
----- ----------- ------------------------ ------------ -----------
 r  b    avm   fre    re  pi  po  fr   sr  cy  in   sy  cs us sy id wa
 0  0 102817 435411   0   0   0   0    0   0   4   12 158  0  0 99  0

  psz   avm   fre   re  pi  po  fr   sr  cy     siz
   4K 84593  189715  0   0   0   0    0   0  293136
  64K  1139  15356   0   0   0   0    0   0   16495
```

in page size units

Multiple Page Sizes - vmstat*(cont)*

```
# vmstat -p 4K,16M 5

System configuration: lcpu=4 mem=2176MB

kthr     memory             page                   faults        cpu
----- ----------- ------------------------ ------------ -----------
 r  b    avm   fre   re   pi   po   fr   sr   cy   in    sy  cs us sy id wa
 0  0 102883 394293    0    0    0    0    0    0    4    33 163  0  0 99  0

  psz    avm    fre    re   pi   po   fr   sr   cy     siz
   4K  84467   169269   0    0    0    0    0    0  272656
  16M      0   10       0    0    0    0    0    0      10

kthr     memory             page                   faults        cpu
----- ----------- ------------------------ ------------ -----------
 r  b    avm   fre   re   pi   po   fr   sr   cy   in    sy  cs us sy id wa
 0  0 102884 394292    0    0    0    0    0    0    6     6 157  0  0 99  0

  psz    avm    fre    re   pi   po   fr   sr   cy     siz
   4K  84468   169268   0    0    0    0    0    0  272656
  16M      0   10       0    0    0    0    0    0      10

kthr     memory             page                   faults        cpu
----- ----------- ------------------------ ------------ -----------
 r  b    avm   fre   re   pi   po   fr   sr   cy   in    sy  cs us sy id wa
 0  0 102884 394292    0    0    0    0    0    0    5     7 156  0  0 99  0

  psz    avm    fre    re   pi   po   fr   sr   cy     siz
   4K  84468   169268   0    0    0    0    0    0  272656
  16M      0   10       0    0    0    0    0    0      10
```

Multiple Page Sizes - vmstat*(cont)*

- Also works in combination with -I

```
# vmstat -I -p ALL 5

System configuration: lcpu=4 mem=2176MB

  kthr     memory             page                  faults        cpu
-------- ----------- ------------------------ ------------ -----------
 r  b  p   avm   fre fi   fo   pi  po  fr  sr  in   sy  cs us sy id wa
 2  0  0 23514 34521 1992 5419 0   0   0   0   0    31  34  0  0 99  0

     psz   avm   fre fi   fo   pi  po  fr  sr siz
      4K 23514 34521 0    0    0   0   0   0 64586

  kthr     memory             page                  faults        cpu
-------- ----------- ------------------------ ------------ -----------
 r  b  p   avm   fre fi   fo   pi  po  fr  sr  in   sy  cs us sy id wa
 1  0  0 23516 34518 1992 5431 0   0   0   0   10    5  40  0  0 96  3

     psz   avm   fre fi   fo   pi  po  fr  sr siz
      4K 23516 34518 0    12   0   0   0   0  64586

# vmstat -I -P ALL 5

System configuration: mem=2176MB

pgsz            memory                           page
----- ------------------------- ----------------------------------
            siz       avm       fre     fi     fo     pi     po     fr      sr
   4K     64586     23516     34518      0      0      0      0      0       0
   4K     64586     23516     34517      0      0      0      0      0       0
```

Multiple Page Sizes - svmon

```
# svmon -G
               size       inuse        free         pin     virtual
memory       557056      162644      394412      106686      102798
pg space     131072         715

               work        pers        clnt
pin             65726           0           0
in use         102798           0       18886

PageSize   PoolSize       inuse        pgsp         pin     virtual
s    4 KB         -      103332         715       57054       84446
m   64 KB         -        1147           0         542        1147
L   16 MB        10           0           0          10           0

# svmon -P 237690
-------------------------------------------------------------------------------
     Pid Command          Inuse       Pin     Pgsp  Virtual 64-bit Mthrd  16MB
  237690 IBM.ServiceRM    17326      8598        0    17224       N     Y     N

    PageSize       Inuse         Pin        Pgsp     Virtual
    s    4 KB      12798        8598           0       12696
    m   64 KB        283           0           0         283
    L   16 MB          0           0           0           0

     Vsid      Esid Type Description              PSize  Inuse   Pin Pgsp Virtual
        0         0 work kernel segment             s  12378  8588    0   12378
   330ad         d work shared library text        m    283     0    0     283
    c2c7         f work shared library data        s    149     0    0     149
    82c6         2 work process private            s    143     3    0     143
   4c297         - clnt /dev/hd9var:287            s     66     0    -       -
   342c9         1 clnt code,/dev/hd2:4977         s     34     0    -       -
   1c2a3         - work                            s     25     7    0      25
   7c2bb         - clnt /dev/hd9var:299            s      1     0    -       -
   302a8         4 work shared memory segment      s      1     0    0       1
    42c5         - clnt /dev/hd9var:297            s      1     0    -       -
   302c8         - clnt /dev/hd9var:286            s      0     0    -       -
   48296         3 mmap maps 1 source(s)           s      0     0    -       -
   54291         - clnt /dev/hd9var:292            s      0     0    -       -
```

in page size units

netpmon - NFS V3 support

- Client report format
  - the number of bytes requested by NFS V3 client read or write RPCs is currently unavailable.
    - the **NFS_RFSRW** hook that has this information when NFS V2 is used is not available for NFS V3.
  - only the number of read and write calls for each client are reported.

```
NFSv3 Server Statistics (by Client):
-------------------------------------

                                    Read            Write            Other
Client                           Calls/s         Calls/s          Calls/s
------------------------------------------------------------------------------
bu.frec.bull.fr                     0.00            0.00             0.21
------------------------------------------------------------------------------
Total (all clients)                 0.00            0.00             0.21

==============================================================================

Detailed NFSv3 Server Statistics (by Client):
-----------------------------------------

CLIENT: bu.frec.bull.fr
other calls:            3
  other times (msec):   avg 0.039   min 0.019   max 0.069   sdev 0.022

COMBINED (All Clients)
other calls:            3
  other times (msec):   avg 0.039   min 0.019   max 0.069   sdev 0.022
```

- All other new reports(by Pid, File and Server) use NFS V2 format

netpmon - NFS V4 support

- Client report format
  - NFS V4 introduced the notion of compound RPCs.
    - a single RPC can contain multiple NFS V4 operations (open, read, write, close ...)
    - compound RPCs are used to improve performance in high latency networks by allowing a client to combine multiple, dependent NFS operations into a single RPC.
  - instead of reporting the number of RPCs, netpmon reports the number of operations issued by NFS V4 clients.

```
 NFSv4 Server Statistics (by Client):
-------------------------------------
                          ------ Read -----    ----- Write -----      Other
Client                     Ops/s    Bytes/s    Ops/s   Bytes/s       Ops/s
-------------------------------------------------------------------------------
bu.frec.bull.fr             0.00          0     0.00         0        0.21
-------------------------------------------------------------------------------
Total (all clients)         0.00          0     0.00         0        0.21

===============================================================================

Detailed NFSv4 Server Statistics (by Client):
---------------------------------------------

CLIENT: bu.frec.bull.fr
other operations:       3
  other times (msec):   avg 0.039   min 0.019   max 0.069   sdev 0.022

COMBINED (All Clients)
other operations:       3
  other times (msec):   avg 0.039   min 0.019   max 0.069   sdev 0.022
```

- All other new reports(by Pid, File and Server) use NFS V2 format, except that they use ops/s instead of calls/s

curt - NFS V4 support

- New NFS V4 subsections in all NFS sections
  - includes new client section

System NFS Calls Summary

Count	Total Time (msec)	Avg Time (msec)	Min Time (msec)	Max Time (msec)	% Tot Time	% Tot Count	Opcode
2015	3056.9121	1.3555	0.1035	31.6976	40.45	17.12	ACCESS
105	2296.3158	15.8367	1.1177	42.9125	22.73	0.82	CLOSE
3025	2263.3336	0.2150	0.0547	2.9737	22.40	59.77	COMMIT
373	777.2854	2.0839	0.2839	17.5724	7.69	2.12	CREATE
2058	385.9510	0.1875	0.0875	1.1993	3.82	11.69	DELEGPURGE
942	178.6442	0.1896	0.0554	1.2320	1.77	5.35	DELEGRETURN
515	97.0297	0.1884	0.0659	0.9774	0.96	2.92	GETATTR
25	11.3046	0.4522	0.2364	0.9712	0.11	0.14	GETFH
3	2.8648	0.9549	0.8939	0.9936	0.03	0.02	LINK
3	2.8590	0.9530	0.5831	1.4095	0.03	0.02	LOCK
2	1.1824	0.5912	0.2796	0.9028	0.01	0.01	LOCKT
1	0.2773	0.2773	0.2773	0.2773	0.00	0.01	LOCKU
1	0.2366	0.2366	0.2366	0.2366	0.00	0.01	LOOKUP
...(lines omitted)...							
17609	10104.3769	0.5738					**NFS V4 SERVER TOTAL**
3	2.8590	0.9530	0.5831	1.4095	0.03	0.02	NFS4_ACCESS
2	1.1824	0.5912	0.2796	0.9028	0.01	0.01	NFS4_VALIDATE_CACHES
1	0.2773	0.2773	0.2773	0.2773	0.00	0.01	NFS4_GETATTR
1	0.2366	0.2366	0.2366	0.2366	0.00	0.01	NFS4_CHECK_ACCESS
1	0.0000	0.0000	0.1804	0.1804	0.00	0.01	NFS4_HOLD
1	0.1704	0.1704	0.1704	0.1704	0.00	0.01	NFS4_RELE
...(lines omitted)...							
17609	10104.3769	0.5738					**NFS V4 CLIENT TOTAL**

tprof update - enhanced shared libraries support

- Privately loaded shared libraries now supported
  - regular routine level breakdown and source annotation available
- New trace -M option to dump process mappings
  - necessary to see private and named shared libraries when using manual offline mode
- Named Shared Libraries
  - previous level of support (introduced in 5.3 ML3) only included breakdown by named area

```
Total % For All Processes (SH-LIBs) = 80%

Total % For All Processes (GLOBAL AREA) = 40.74%

Shared Object                                                    %
=============                                                ======
/usr/lib/libc.a[shr.o]                                        40.74

   Profile: /usr/lib/libc.a[shr.o]

   Total % For All Processes (/usr/lib/libc.a[shr.o]) = 40.74

Subroutine                                                       %   Source
==========                                                   ====== ======
._doprnt                                                      28.53 cs/lib/libc/doprnt.c
.strlen                                                        4.80 strlen.s
.printf                                                        4.53 cs/lib/libc/printf.c
.strchr                                                        2.88 strchr.s

Total % For All Processes (foo AREA) = 10%
Total % For All Processes (test AREA) = 20%
Total % For All Processes (bar AREA) = 9.26%
```

tprof update - full named shared libraries support

- Includes routine breakdown

```
Configuration information
=========================
System: AIX 5.3 Node: gibi Machine: 00CFEDAD4C00
Tprof command was:
    tprof -suz -t -r report -x LDR_CNTRL=NAMEDSHLIB=zone1 vloop_lib_32 5
Trace command was:
    /usr/bin/trace -ad -M -L 203408179 -T 500000 -j 000,00A,001,002,003,38F,005,006,134,139,5A2,5A5,465,234, -o -

...


  Total % For ./vloop_lib_32[254068] thread 614471 (SH-LIBs zone1) = 31.30

Shared Object                                                    %
=============                                                ======
/usr/lib/libc.a[shr.o]                                        27.14
./libloop.a[lloop_lib_32]                                      4.16

  Profile: /usr/lib/libc.a[shr.o]

  Total % For ./vloop_lib_32[254068] thread 614471 (/usr/lib/libc.a[shr.o]) = 27.14

Subroutine                                          %    Source
==========                                        ====== ======
.free_y                                            10.27 /lib/libc/malloc_y.c
.malloc_y                                          10.02 /lib/libc/malloc_y.c
.free_common                                        1.71 libc/malloc_common.c
.leftmost                                           1.47 /lib/libc/malloc_y.c
.heap_select_y                                      0.73 /lib/libc/malloc_y.c
.splay                                              0.73 /lib/libc/malloc_y.c
._ptrgl                                             0.73 ptrgl.s
.malloc_common                                      0.73 libc/malloc_common.c
.srand                                              0.49 /ccs/lib/libc/rand.c
._ptrgl                                             0.24 ptrgl.s

  Profile: ./libloop.a[lloop_lib_32]

  Total % For ./vloop_lib_32[254068] thread 614471 (./libloop.a[lloop_lib_32]) = 4.16

Subroutine                                          %    Source
==========                                        ====== ======
.lloop                                              3.18 lloop.c
.free                                               0.49 glink.s
.malloc                                             0.49 glink.s
```

tprof enhancement - privately loaded libraries

- Example in combination with named shared library usage

```
  Total Ticks For ./vloop_lib_32[254134] thread 569451 (SH-LIBs zone1) = 130

Shared Object                                    Ticks     %    Address  Bytes
=============                                    ===== ======  =======  =====
/usr/lib/libc.a[shr.o]                             111  22.98 d01083a0 22877c
./libloop.a[lloop_lib_32]                           19   3.93 2000014c    190

  Profile: /usr/lib/libc.a[shr.o]

  Total Ticks For ./vloop_lib_32[254134] thread 569451 (/usr/lib/libc.a[shr.o]) = 111

Subroutine                 Ticks     %   Source                Address  Bytes
==========                 ===== ====== ======                =======  =====
.free_y                       45   9.32 /lib/libc/malloc_y.c     c694    6dc
.malloc_y                     39   8.07 /lib/libc/malloc_y.c     da20    604
._ptrgl                        5   1.04 ptrgl.s                   250      1
._ptrgl                        5   1.04 ptrgl.s                   250     30
.splay                         4   0.83 /lib/libc/malloc_y.c     bc00    3b4
.malloc_common_53_36           3   0.62 libc/malloc_common.c     b4f8     58
.free_common                   3   0.62 libc/malloc_common.c     ad30     c0
.malloc_common                 3   0.62 libc/malloc_common.c     af5c     38
.leftmost                      2   0.41 /lib/libc/malloc_y.c     c0dc    124
.srand                         2   0.41 /ccs/lib/libc/rand.c   1320fc     90

  Profile: ./libloop.a[lloop_lib_32]

  Total Ticks For ./vloop_lib_32[254134] thread 569451 (./libloop.a[lloop_lib_32] private) = 19

Subroutine                 Ticks     %   Source                Address  Bytes
==========                 ===== ====== ======                =======  =====
.lloop                        17   3.52 lloop.c                    50    140
.free                          1   0.21 glink.s                    28     28
.free                          1   0.21 glink.s                    28      1
```

tprof update - hypervisor support

- New category reporting total time spent in hypervisor
  - only available when -E is used
- Represents time spent servicing hcalls
  - does not include full hardware context switching time
  - no routine breakdown available
    - offset level breakdown available when -D is selected

```
Process                          Freq  Total Kernel   User Shared  Other
=======                          ====  ===== ======   ==== ======  =====
/usr/bin/yes                        1  35.02   0.40   1.30  33.32   0.00
./vloop_lib_32                      1  34.79   0.18   3.39  31.22   0.00
wait                                4  30.01  30.01   0.00   0.00   0.00
/usr/bin/tprof                      2   0.09   0.09   0.00   0.00   0.00
/usr/bin/sh                         1   0.04   0.00   0.04   0.00   0.00
/usr/bin/trcstop                    1   0.04   0.04   0.00   0.00   0.00
=======                          ====  ===== ======   ==== ======  =====
Total                              10 100.00  30.73   4.73  64.54   0.00

...

  Total % For All Processes (KERNEL) = 1.16

Subroutine                                    %   Source
==========                                ======  ======
.kdb_state_restore_to_pr_flih                0.27  64/low.s
_tls_                                        0.18  64/low.s
state_save_fixup                             0.04  state_sslb.s

...

  Total % For All Processes (HYPERVISOR) = 29.57

Subroutine                                    %   Source
==========                                ======  ======
<0x8000000000E05BC8>                        10.41
<0x8000000000E05BC4>                         9.51
...
```

tprof update - millicode support

- Millicode is now reported separatedly in **two** new sections
  - in shared library section for user mode calls

```
  Total % For All Processes (SH-LIBs) = 40.66

Shared Object                                                    %
=============                                                ======
/usr/lib/libc.a[shr.o]                                        39.56
Millicode routines                                             1.10

  Profile: /usr/lib/libc.a[shr.o]

  Total % For All Processes (/usr/lib/libc.a[shr.o]) = 39.56

Subroutine                                          %    Source
==========                                       ====== ======
._doprnt                                          21.98 cs/lib/libc/doprnt.c
.strlen                                           10.99 strlen.s
.printf                                            4.40 cs/lib/libc/printf.c
.strchr                                            2.20 strchr.s

  Profile: Millicode routines

  Total % For All Processes (Millicode routines) = 1.10

Subroutine                                          %    Source
==========                                       ====== ======
.mulh                                              1.10 64/low.s
```

  - in kernel section for kernel mode calls

```
Total % For All Processes (KERNEL) = 50.50

Subroutine                                          %    Source
==========                                       ====== ======
h_cede_end_point                                  46.76 hcalls.s
.waitproc_find_run_queue                           0.50 rnel/proc/dispatch.c
.ufdrele                                           0.25 /bos/kernel/lfs/fd.c
.v_pagein                                          0.25 nel/vmm/v_getsubs1.c
...

Milicode Subroutine                                 %    Source
===================                              ====== ======
.mulh                                             1.25 64/low.s
```

Hardware PM updates

- Power5+ support (in 5.3 ML 4 and 5.2 ML 8)
  - 183 groups (Power5 had 148)
    - most Power5 groups still exist, but do not always have the same number
  - pmc5 is now counting PM_RUN_INST_CMPL

- New set of APIs reporting time
  - pm_tstart* and pm_tstop*
    - return timestamps(time base values) when counting started or stopped
    - can be used in combination with existing **pm_get_tdata*** interfaces to measure counting intervals
  - pm_get_Tdata*
    - report measurement interval in TB, PURR and SPURR units, e.g.

```
typedef struct {
     timebasestruct_t accu_timebase;  /* accumulated time base  */
     timebasestruct_t accu_purr;      /* accumulated PURR time  */
     timebasestruct_t accu_spurr;     /* accumulated SPURR time */
  } pm_accu_time_t;

pm_get_Tdata(pm_data_t *data, pm_accu_time_t *times);
```

- Counter multiplexing
  - ability to count more events than number of physical counters
  - supported by libpmapi, libhpm, hpmcount and hpmstat
    - new set of **pm_*_mx** interfaces
    - expanded command line syntax for hpmcount and hpmstat to support multiple event sets
    - expanded syntax for libhpm/hpmcount/hpmstat environment variables to support multiple event sets

PMAPI update - counter multiplexing

- New data structures

```
typedef int pm_events_prog_t[MAX_COUNTERS];
typedef struct {                                /* structure for PM programing          */
  pm_mode_t        mode;                        /* mode of operation                    */
  int              slice_duration;              /* duration of each time slice in ms    */
  int              nb_events_prog;              /* number of events_set                 */
  pm_events_prog_t *events_set;                 /* list of counted events               */
} pm_prog_mx_t;

typedef struct {
  timebasestruct_t accu_time;                   /* accumulated time                     */
  timebasestruct_t accu_purr;                   /* accumulated PURR time                */
  timebasestruct_t accu_spurr;                  /* accumulated SPURR time               */
  long long        accu_data[MAX_COUNTERS];     /* accumulated data                     */
} pm_accu_mx_t;

typedef struct {                                /* structure for PM data                */
  pm_ginfo_t   ginfo;                           /* group information                    */
  int          nb_accu_mx;                      /* number of accu_set                   */
  int          nb_mx_round;                     /* number of loops on all the event sets */
  pm_accu_mx_t *accu_set;                       /* accumulated data                     */
} pm_data_mx_t;
```

- Example of new interfaces

```
int     pm_set_program_mx(pm_prog_mx_t *prog)     [compares to pm_set_program(pm_prog_t *prog)]
int     pm_get_program_mx(pm_prog_mx_t *prog)     [compares to pm_get_program(pm_prog_t *prog)]
int     pm_get_data_mx(pm_data_mx_t *data)        [compares to pm_get_data(pm_data_t *data)]
```

- hpmcount and hpmstat support
  - -s flag now allows comma separated list of event sets to be specified
    - set "0" means all sets
  - environment variables similarly now accepts multiple comma separated sets
  - multiple groups can be specified via event file

hpmcount - example of multiplexing all sets

```
# hpmcount -s 0 ipc4
 Execution time (wall clock time): 64.697222 seconds

 ########  Resource Usage Statistics  ########

 Total amount of time in user mode            : 64.339401 seconds
 Total amount of time in system mode          : 0.017005 seconds
 Maximum resident set size                    : 388 Kbytes
 Average shared memory use in text segment    : 257 Kbytes*sec
 Average unshared memory use in data segment  : 24757 Kbytes*sec
 Number of page faults without I/O activity   : 140
 Number of page faults with I/O activity      : 0
 Number of times process was swapped out      : 0
 Number of times file system performed INPUT  : 0
 Number of times file system performed OUTPUT : 0
 Number of IPC messages sent                  : 0
 Number of IPC messages received              : 0
 Number of signals delivered                  : 0
 Number of voluntary context switches         : 2
 Number of involuntary context switches       : 6656

 #######  End of Resource Statistics  ########

  PM_LSU_CMPL (LSU instructions completed)                                  :      7981013360
  PM_CYC (Processor cycles)                                                 :     24001739529
  PM_INST_CMPL (Instructions completed)                                     :     32000866113
  PM_INST_DISP (Instructions dispatched)                                    :     31992690593
  PM_IC_MISS (Instruction cache misses)                                     :            8068
  PM_LSU_IDLE (Cycles LSU is idle)                                          :     16006473444
  PM_SNOOP (Snoop requests received)                                        :           29310
  PM_SNOOP_HIT (Snoop hits)                                                 :               8
  PM_FPU_CMPL (Floating-point instructions completed (no loads or stores))  :               0
  PM_FXU_CMPL (Integer instructions completed (no loads or stores))         :     16007417946
  PM_DTLB_MISS (Data TLB misses)                                            :             674
  PM_ITLB_MISS (Instruction TLB misses)                                     :             134
  PM_BR_MPRED (Branches incorrectly predicted)                              :               0
  PM_BR_DISP (Instructions dispatched to the branch unit)                   :      8004870010


  Processing time                                 :          64.005 s
  Utilization rate                                :          98.930 %
  Instructions per cycle                          :           1.333
  MIPS                                            :         494.625 MIPS
  % Instructions dispatched that completed        :         100.026 %
  Total load and store operations                 :        7981.013 M
  Instructions per load/store                     :           4.010
  Instructions per I Cache Miss                   :           4.010
  % Cycles LSU is idle                            :          66.689 %
  Snoop hit rate                                  :           0.027 %
  HW floating point instructions per Cycle        :           0.000
  HW floating point instructions / user time      :           0.000 M HWflops/s
  HW floating point rate                          :           0.000 M HWflops/s
  Total Fixed point operations                    :       16007.418 M
  Fixed point operations per Cycle                :           0.667
  Branches mispredicated percentage               :           0.000 %
```

PMAPI - Dynamic Reconfiguration support

- Processor additions and deletion now supported
  - includes turning SMT on or off
- Impact to per-cpu interfaces
  - pm_get_data_cpu, pm_get_tdata_cpu and the new pm_get_Tdata_cpu and pm_get_data_cpu_mx
    - cpuids are always contiguous (0 to _systemcfg.**ncpus**)
    - may not always represent the same logical processors
    - DR operations renumber cpus
    - partial results for deleted cpus are lost
  - new pm_get_data_lcpu and pm_get_data_lcpu_mx interfaces
    - lcpuids are not always contiguous (0 to __systemcfg.**max_ncpus**)
    - always represent the same logical processor
    - DR operations create or fill holes in lcpuids
    - partial results for deleted cpus can be retrieved

topas - CEC monitoring screen(5.3 ML 3)

- Split screen accessible with -C or the "C" command
  - Upper section shows CEC-level metrics
  - Lower sections shows sorted list of shared and dedicated partitions

```
Topas CEC Monitor            Interval:  10              Thu Jul 28 17:04:57 2005
Partitions     Memory (GB)                  Processors
Shr:  3        Mon:24.6  InUse: 2.7         Shr:1.5  PSz:  3   Shr_PhysB: 0.27
Ded:  3        Avl:   -                     Ded:  5  APP:  2.6 Ded_PhysB: 2.70

Host         OS  M Mem InU Lp  Us Sy Wa Id  PhysB  Ent  %EntC Vcsw PhI
-----------------------------------shared-----------------------------------
ptoolsl3     A53 c 4.1 0.4  2  14  1  0 84  0.08  0.50  15.0  208   0
ptoolsl2     A53 C 4.1 0.4  4  20 13  5 62  0.17  0.50  36.5  219   5
ptoolsl5     A53 U 4.1 0.4  4   0  0  0 99  0.02  0.50   0.1  205   2

----------------------------------dedicated---------------------------------
ptoolsl1     A53 S 4.1 0.5  4  20 10  0 70  0.60
ptoolsl4     A53   4.1 0.5  2 100  0  0  0  2.00
ptoolsl6     A52   4.1 0.5  1   5  5 12 88  0.10
```

- Configuration info retrieved from HMC or specified from command line
  - c means capped, C - capped with SMT
  - u means shared, U - uncapped with SMT
  - S means SMT
- Uses new **xmtopas** daemon started by inetd

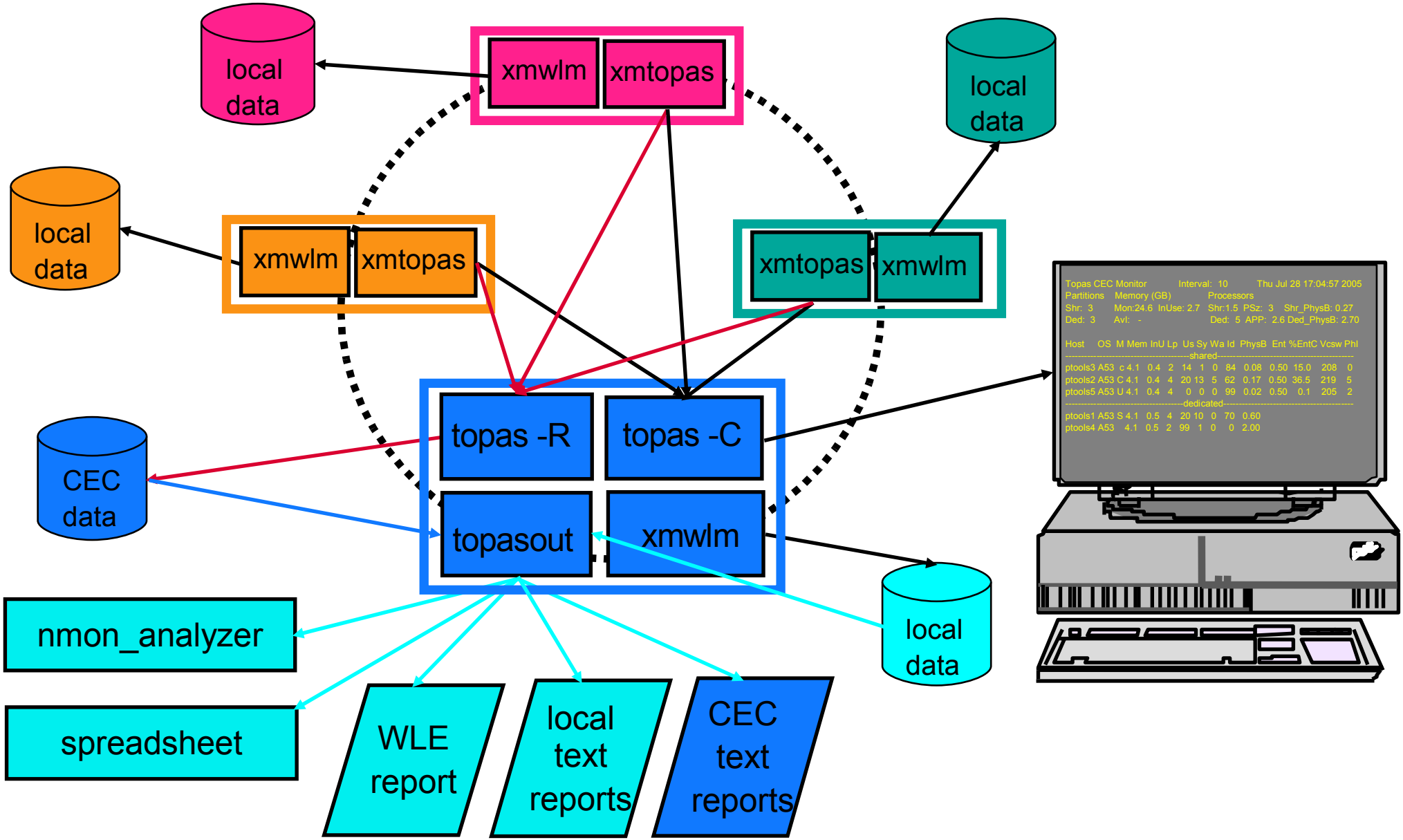
Automatic Performance Metric recording

- Introduced in 5.3 ML 4
  - uses xmwlm daemon
  - automaticaly started from inittab
  - initially kept 2 days worth of data, but changing to 7 days in 5.3 TL5
  - recordings include most of topas data
    - except process and WLM data
- New (5.3 TL5) topas -R option records topas -C metrics (CEC-wide data)
  - works independently and in parallel from topas real-time monitors
  - must be turned on manually in one of the partitions in CEC
    - via configuration script which adds line in inittab

```
# /usr/lpp/perfagent/config_topas.sh add
```

- topasout
  - postprocessing tool for recordings
  - WLE reports
  - text reports (5.3 TL5)
    - include both local data and CEC-wide data
    - options include detailed and summary
  - spreadsheet and csv formats
  - nmon_analyzer format (5.3 TL5)

Automatic Performance Metric recording*(cont)*



topasout - CEC summary report

- Example with configuration change

```
Report: Topas CEC Summary --- hostname: ptoolsl1                    version:1.0
Start:02/09/06 00.00.00 Stop:02/09/06 23.55.00  Int:  5 Min  Range:1440 Min
Partition Mon:  7  UnM:  1  Shr:  4   Ded:  3   Cap:  3   UnC:  1
          -CEC------  -Processors-------------------- -Memory (GB)------------
Time      ShrB  DedB  Mon UnM Avl UnA Shr Ded PSz APP  Mon  UnM  Avl  UnA  InU
00.05.00   3.2   1.1    5   2   7   1   4   3   2   1  16.0 0.0  32.0 0.0  8.1
00.10.00   2.9   0.9    5   2   7   1   4   3   2   1  16.0 0.0  32.0 0.0  8.3
00.15.00   2.1   1.3    5   2   7   1   4   3   2   1  16.0 0.0  32.0 0.0  8.5

...

configuration change at 02.15.00
Partition Mon:   8  UnM:  0  Shr:  4   Ded:  4   Cap:  3   UnC:  1
          -CEC------  -Processors-------------------- -Memory (GB)------------
Time      ShrB  DedB  Mon UnM Avl UnA Shr Ded PSz APP  Mon  UnM  Avl  UnA  InU
02.15.00   3.1   2.5    7   0   7   1   4   5   2   1  18.0 0.0  32.0 0.0  9.1
02.20.00   1.9   1.5    7   0   7   1   4   5   2   1  18.0 0.0  32.0 0.0  6.8
02.25.00   2.0   3.3    7   0   7   1   4   5   2   1  18.0 0.0  32.0 0.0  7.8

...
```

topasout - detailed CEC report

```
Report: Topas CEC Detailed --- hostname: ptoolsl1                  version:1.0
Start:05/02/06 07.00.00  Stop:05/02/06 17.00.00  Int:05 Min  Range:600 Min
Time: 07.00.00 ----------------------------------------------------------------
Partition Info     Memory (GB)          Processors
Monitored   :  8   Monitored   : 0.0   Monitored   :  7   Shr Physical Busy:  2.2
UnMonitored:  -   UnMonitored: 0.0   UnMonitored:  0   Ded Physical Busy:  0.4
Shared      :  3   Available   :32.0   Available   :  7
Dedicated   :  2   UnAllocated:   -   UnAllocated:  1   Hypervisor
Capped      :  1   Consumed    : 8.7   Shared      :  4   Virt. Context Switch:332
Uncapped    :  2                       Dedicated   :  3   Phantom Interrupts  :  2
                                       Pool Size   :  2
                                       Avail Pool  :  1
Host          OS  M Mem InU Lp  Us Sy Wa Id  PhysB  Ent  %EntC Vcsw PhI
-------------------------------shared-------------------------------------------
ptools1       A53 u 1.1 0.4  4  15  3  0 82   1.30  0.50  22.0  200   5
ptools5       A53 U  12  10  1  12  3  0 85   0.20  0.25   0.3  121   3
ptools3       A53 C 5.0 2.6  1  10  1  0 89   0.15  0.25   0.3  52    2
-------------------------------dedicated----------------------------------------
ptools4       A53 S 0.6 0.3  2  12  3  0 85   0.60
ptools6       A52   1.1 0.1  1  11  7  0 82   0.50
ptools8       A52   1.1 0.1  1  11  7  0 82   0.50
Time: 07.00.05 ----------------------------------------------------------------
Partition Info     Memory (GB)          Processors
Monitored   :  8   Monitored   : 0.0   Monitored   :  7   Shr Physical Busy:  2.2
UnMonitored:  -   UnMonitored: 0.0   UnMonitored:  0   Ded Physical Busy:  0.4
Shared      :  3   Available   :32.0   Available   :  7
Dedicated   :  2   UnAllocated:   -   UnAllocated:  1   Hypervisor
Capped      :  2   Consumed    : 8.7   Shared      :  4   Virt. Context Switch:332
Uncapped    :  2                       Dedicated   :  3   Phantom Interrupts  :  2
                                       Pool Size   :  2
                                       Avail Pool  :  1
Host          OS  M Mem InU Lp  Us Sy Wa Id  PhysB  Ent  %EntC Vcsw PhI
-------------------------------shared-------------------------------------------
ptools1       A53 u 1.1 0.4  4  15  3  0 82   1.30  0.50  22.0  200   5
ptools5       A53 U  12  10  1  12  3  0 85   0.20  0.25   0.3  121   3
ptools3       A53 C 5.0 2.6  1  10  1  0 89   0.15  0.25   0.3  52    2
-------------------------------dedicated----------------------------------------
ptools4       A53 S 0.6 0.3  2  12  3  0 85   0.60
ptools6       A52   1.1 0.1  1  11  7  0 82   0.50
ptools8       A52   1.1 0.1  1  11  7  0 82   0.50
Time: 07.00.10 ----------------------------------------------------------------
```

topasout - summary local report

- Dedicated partitions

```
Report: System Summary - hostname: ptoolsl1                    version 1.0

Start:12/20/05 14.00.00 Stop:12/20/05 15.00.00  Int: 5 Min  Range: 60 Min
Mem: 16.2 GB  Dedicated SMT:OFF  Logical CPUs:  2
Time      InU  Us Sy Wa Id  PhysB  RunQ WtQ CSwitch Syscall PgFault
-------------------------------------------------------------------------
14.00.00  21.1 11  8  0 81    0.2     1   0    3432    5050      17
14.05.00  21.1 16  5  0 79    0.3     1   0     532    3104      14
14.10.00  21.2 13  7  0 20    0.2     1   0     652    4326      13
```

- Shared partitions

```
Report: System Summary - hostname: ptoolsl1                    version 1.0

Start:12/21/05 10.00.00 Stop:12/21/05 11.00.00  Int: 5 Min  Range: 60 Min
Psize:1.0  Mem: 16.2 GB  Shared SMT:OFF  Logical CPUs:  2
Time      InU  Us Sy Wa Id  PhysB  Ent  %EntC RunQ WtQ CSwitch Syscall PgFault
-------------------------------------------------------------------------------
10.00.00  21.1 11  8  0 81    0.2  0.5   23.2    1   0    3432    5050      17
10.05.00  21.1 16  5  0 79    0.3  0.5   25.0    1   0     532    3104      14
10.10.00  21.2 13  7  0 20    0.2  0.5   23.4    1   0     652    4326      13
```

topasout - detailed local report

```
Report: System Detailed --- hostname: ptools11                    version 1.0
Start:12/21/05 10.00.00 Stop:12/21/05 11.00.00  Int: 5 Min  Range: 60 Min
Time: 10.00.00 ----------------------------------------------------------------
CPU           UTIL          MEMORY          PAGING          EVENTS/QUEUES  NFS
Kern  12.0  PhyB   0.7  Sz,GB  16.0  Sz,GB  4.0  Cswth    3213  SrvV2    32
User   8.0  Ent    0.5  InU     4.3  InU    2.3  Syscl   43831  CltV2    12
Wait   0.0  EntC  15.2  %Comp   3.1  Flt    221  RunQ        1  SrvV3    44
Idle  78.0  LP       4  %NonC   9.0  Pg-I    87  WtQ         0  CltV3    18
SMT     ON  Mode   Shr  %Clnt   2.0  Pg-O    44  VCSW     1214

Network   KBPS    I-Pack   O-Pack     KB-I     KB-O
en0        0.6       7.5      0.5      0.3      0.3
en1       22.3     820.1    124.3    410.0     61.2
lo0        0.0       0.0      0.0      0.0      0.0

Disk     Busy%      KBPS      TPS     KB-R     KB-W
hdisk0     0.0       0.0      0.0      0.0      0.0
hdisk1     0.0       0.0      0.0      0.0      0.0
Time: 10.05.00 ----------------------------------------------------------------
CPU           UTIL          MEMORY          PAGING          EVENTS/QUEUES  NFS
Kern  12.0  PhyB   0.7  Sz,GB  16.0  Sz,GB  4.0  Cswth    3213  SrvV2    32
User   8.0  Ent    0.5  InU     4.3  InU    2.3  Syscl   43831  CltV2    12
Wait   0.0  EntC  15.2  %Comp   3.1  Flt    221  RunQ        1  SrvV3    44
Idle  78.0  LP       4  %NonC   9.0  Pg-I    87  WtQ         0  CltV3    18
SMT     ON  Mode   Shr  %Clnt   2.0  Pg-O    44  VCSW     1214

Network   KBPS    I-Pack   O-Pack     KB-I     KB-O
en0        0.6       7.5      0.5      0.3      0.3
en1       22.3     820.1    124.3    410.0     61.2
lo0        0.0       0.0      0.0      0.0      0.0

Disk     Busy%      KBPS      TPS     KB-R     KB-W
hdisk0     0.0       0.0      0.0      0.0      0.0
hdisk1     0.0       0.0      0.0      0.0      0.0
```

topasout - I/O summary reports

- Disk report

```
Report: Total Disk I/O Summary - hostname: ptoolsl1                  version:1.0
Start:04/25/06 00.00.00  Stop:04/26/06 00.00.00  Int:05 Min  Range:1440 Min
Mem: 8.0 GB  Dedicated SMT:ON  Logical CPUs:16
Time        InU    PhysB   %Bsy   MBPS    TPS   MB-R   MB-W
-------------------------------------------------------------------------------
00.00.05    6.5    12.50   45.5 120.5 300.1 100.1   20.4
00.00.10    6.7    13.40   55.0 240.0 320.2 240.0    0.0
00.00.15    7.0    14.70   60.4 160.2 350.3  40.1 120.1
00.00.20    7.4    15.50   72.3 200.7 410.5  20.3 180.4
```

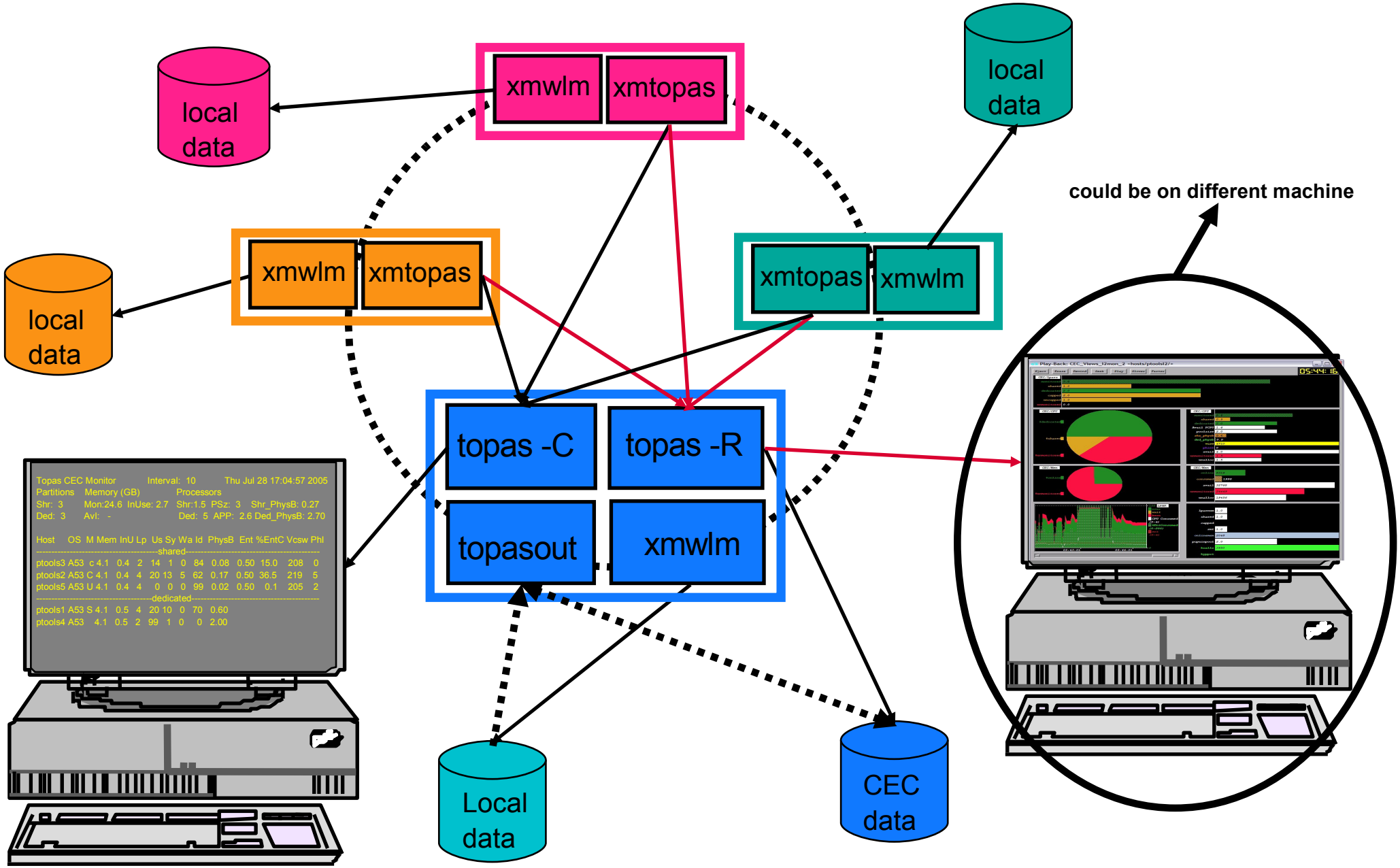
- LAN report

```
Report: Total LAN I/O Summary - hostname: ptoolsl1                   version:1.0
Start:03/12/06 17.15.00  Stop:03/12/06 20.30.00  Int:05 Min  Range: 195 Min
Psize:1.0  Mem: 16.2 GB  Shared SMT:OFF  Logical CPUs:  2
Time        InU    PhysB    MBPS   I-Pack O-Pack MB-I   MB-O Rcvdrp Xmtdrp
-------------------------------------------------------------------------------
17.15.00    3.2     6.30    20.0   310.5  120.2  16.2    3.8 120     160
17.20.00    3.3     6.45    22.3   220.3  225.7  11.1   11.2 118     165
17.25.00    3.2     6.15    18.5   275.6  158.0  11.6    6.9 121     162
17.30.00    3.4     6.55    19.4   270.2  156.9  11.3    6.1 124     154
```

PTX refresh - enhanced virtualization support

- Adds sample consoles for easy physical machine monitoring
  - skeleton console to display real-time aggregated data
    - uses topas CEC recording capability
      - topas -R exports aggregated metrics to PTX name space(needs APAR IY87433)
    - only needs hostname of partition running topas -R to instantiate
  - skeleton consoles for
    - 5.2 partitions
    - 5.3 dedicated partitions
    - 5.3 capped partitions
    - 5.3 uncapped partitions
    - only needs partitions hostnames to instantiate
- Provides easily customizable solution
  - only list of partitions hostnames is needed to instantiate fully functional physical machine monitoring set of consoles
  - includes standard PTX attributes
    - recordings/playback
    - reports

PTX aggregated metrics monitoring

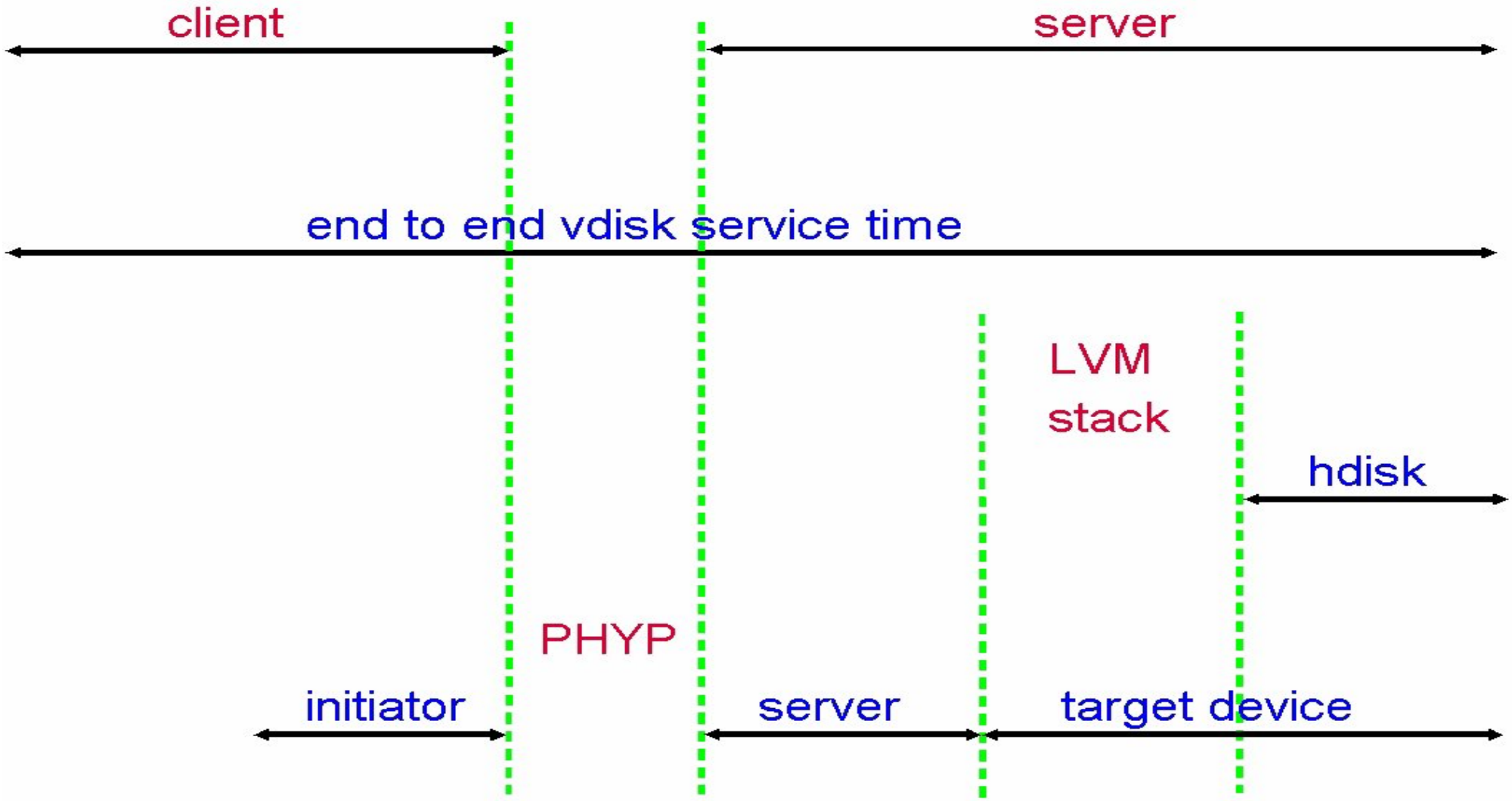


PTX aggregated metrics viewing console

VIOS monitoring

- Current tools available on server
  - topas
    - main screen and -D (detailed disk stats) screnn
  - entstat
  - viostat
    - wrapper around iostat
- Current tools available on client
  - same as server
    - except real iostat instead of viostat
  - lparstat, mpstat, vmstat, sar

Virtual Disk I/O Monitoring - instrumentation



Virtual Disk client monitoring

```
# iostat -a -D

System configuration: lcpu=2 drives=3 paths=1 vdisks=1

Adapter:
scsi0             xfer:       bps       tps     bread     bwrtn
                              0.0       0.0       0.0       0.0
Paths/Disk:
hdisk0_path0      xfer:   %tm_act       bps       tps     bread     bwrtn
                              0.0       0.0       0.0       0.0       0.0
                  read:       rps   avgserv   minserv   maxserv   timeouts      fails
                              0.0       0.0       0.0       0.0          0          0
                 write:       wps   avgserv   minserv   maxserv   timeouts      fails
                              0.0       0.0       0.0       0.0          0          0
                 queue:   avgtime   mintime   maxtime   avgwqsz    avgsqsz    sqfull
                              0.0       0.0       0.0       0.0        0.0          0

Vadapter
vsci0             xfer:       tps     bread     bwrtn  partition-id
                              0.0       0.0       0.0      ####
                  read:   avgserv   minserv   maxserv
                              0.0       0.0       0.0
                 write:   avgserv   minserv   maxserv
                              0.0       0.0       0.0
                 queue:   avgtime   mintime   maxtime   avgsqsz    qfull
                              0.0       0.0       0.0       0.0        0
Disk:
hdisk10           xfer:   %tm_act       bps       tps     bread     bwrtn
                              0.0       0.0       0.0       0.0       0.0
                  read:       rps   avgserv   minserv   maxserv   timeouts      fails
                              0.0       0.0       0.0       0.0          0          0
                 write:       wps   avgserv   minserv   maxserv   timeouts      fails
                              0.0       0.0       0.0       0.0          0          0
                 queue:   avgtime   mintime   maxtime   avgwqsz    avgsqsz    sqfull
                              0.0       0.0       0.0       0.0        0.0          0
```

Virtual Disk server monitoring

```
# viostat -adapter -extdisk

System configuration: lcpu=2 drives=3 paths=1 vdisks=1

Adapter:
scsi0          xfer:       bps       tps     bread     bwrtn
                            0.0       0.0       0.0       0.0
Paths/Disk:
hdisk0_path0   xfer:   %tm_act       bps       tps     bread     bwrtn
                           0.0       0.0       0.0       0.0       0.0
               read:       rps   avgserv   minserv   maxserv   timeouts    fails
                           0.0       0.0       0.0       0.0          0        0
              write:       wps   avgserv   minserv   maxserv   timeouts    fails
                           0.0       0.0       0.0       0.0          0        0
              queue:   avgtime   mintime   maxtime   avgwqsz    avgsqsz   sqfull
                           0.0       0.0       0.0       0.0       0.0        0

Vadapter
vhost0         xfer:       tps     bread     bwrtn
                           0.0       0.0       0.0
               read:   avgserv   minserv   maxserv
                           0.0       0.0       0.0
              write:   avgserv   minserv   maxserv
                           0.0       0.0       0.0
              queue:   avgtime   mintime   maxtime   avgsqsz    qfull
                           0.0       0.0       0.0       0.0        0
Vtarget/Disk:
hdisk0_vtscsi0 xfer:   %tm_act       bps       tps     bread     bwrtn
                           0.0       0.0       0.0       0.0       0.0
               read:       rps   avgserv   minserv   maxserv   timeouts    fails
                           0.0       0.0       0.0       0.0          0        0
              write:       wps   avgserv   minserv   maxserv   timeouts    fails
                           0.0       0.0       0.0       0.0          0        0
              queue:   avgtime   mintime   maxtime   avgwqsz    avgsqsz   sqfull
                           0.0       0.0       0.0       0.0       0.0        0

#lv00_vtscsi1 xfer:   %tm_act       bps       tps     bread     bwrtn
                           0.0       0.0       0.0       0.0       0.0
               read:       rps   avgserv   minserv   maxserv   timeouts    fails
                           0.0       0.0       0.0       0.0          0        0
              write:       wps   avgserv   minserv   maxserv   timeouts    fails
                           0.0       0.0       0.0       0.0          0        0
              queue:   avgtime   mintime   maxtime   avgwqsz    avgsqsz   sqfull
                           0.0       0.0       0.0       0.0       0.0        0
```

VIOS Monitoring - planned enhancements

- 2006
  - xmtopas
    - makes VIOS partitions visible to topas -C and -R
- 2007
  - xmwlm
    - automatically recording of all statistics displayed by topas (except -C data)
  - topasout
    - report generagtor for all recordings

Thank You!