Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Instrumentation

BSC Performance Tools

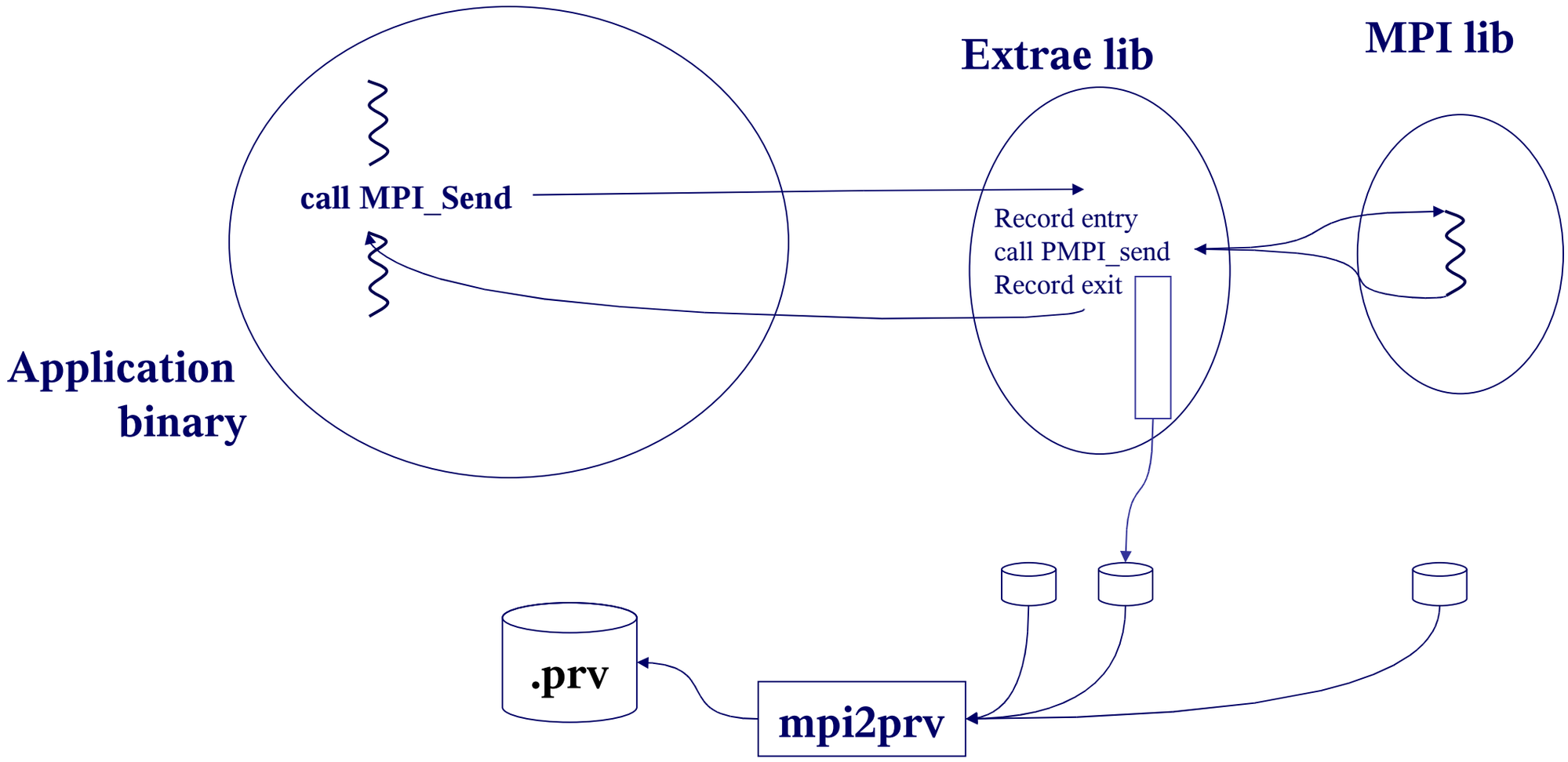
Index

- The instrumentation process
- A typical MN process
- Paraver trace format
- Configuration
  - XML
  - Environment variables
- Adding references to the source
  - API

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

MPItrace: Tracing internals

- Dynamic library calls instrumentation



Parallel Program Instrumentation: Features

- Probe injection mechanisms
  - Library Preload: Linux clusters (Dynamically linked MPI library)
  - Static linking: BG
  - Compiler-based instrumentation (PDT) ALTIX, PowerPC
  - dynamic instrumentation (Dyninst): ALTIX, PowerPC (beta version)
- Captured events:
  - MPI calls (including I/O)
  - Hardware counters: PAPI: standard + native
    - Several sets: Rotate groups periodically / Different groups per processes
  - Network: GM at end of trace, MX at flushes
  - OS counters: At the end of the trace and when flushing
  - Link to source:
    - Dyninst based systems: user function events on entry and exit (for selected functions).
    - Library preload: MPI caller (several levels)
  - User events (API)
- Towards a unified tracing package and specification
  - xml control file

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

A typical MareNostrum process

mn_tracerun

```
#!/bin/bash
…
# @ total_tasks = 2
# @ cpus_per_task = 1
# @ tasks_per_node = 2
…
srun ./trace.sh ./mpi_ping
```

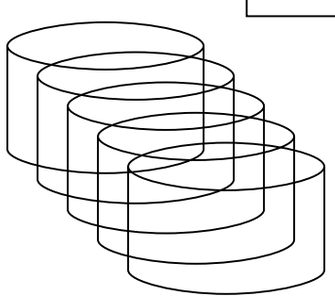
trace.sh

```
export#!/bin/sh

export EXTRAE_HOME=/gpfs/apps/CEPBATOOLS/beta/extrae-sampling/64
export EXTRAE_CONFIG_FILE=extrae.xml

export LD_PRELOAD=${EXTRAE_HOME}/lib/libmpitrace.so

## Run the desired program
$*
```



TRACExxxxxx.mpit

TRACE.mpits

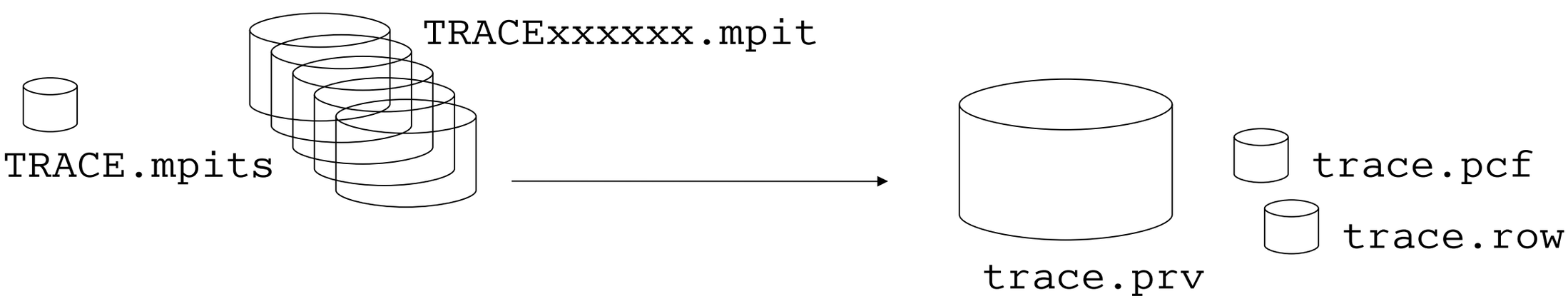
Examples of MN set-up available at /gpfs/apps/CEPBATOOLS/tracing-setup
(using previous version named MPItrace)

A typical MareNostrum process

mn_parmerge

```
#!/bin/bash
…
# @ total_tasks = 4
# @ cpus_per_task = 1
# @ tasks_per_node = 4
…
EXTRAE_HOME=/gpfs/apps/CEPBATOOLS/beta/extrae-sampling/64.hwc

srun ${MPITRACE_HOME}/bin/mpimpi2prv -syn -f TRACE.mpits -o trace.prv
```



Trace records (.prv)

```
#Paraver (10/07/02 at 09:55):1149147987_ns:0:1:4(1:0,1:0,1:0,1:0),2
C:1:0:4:1:2:3:4
C:1:1:1:-1

2:0:1:4:1:0:40000001:1
1:0:1:1:1:0:6771087:2
1:0:1:2:1:0:18830180:2
1:0:1:3:1:0:12803673:2
1:0:1:4:1:0:80823:1
2:0:1:4:1:80823:50000003:31
2:0:1:4:1:80823::42000003:11918
2:0:1:4:1:80823:42000121:141
…
1:0:1:4:1:80823:584006:15
…
…
…
1:0:1:1:1:11331439:11404156:1
3:0:1:1:1:11404156:11687636:0:1:2:1:40124416:40337085:100:0
2:0:1:1:1:11404156:50000001:1:42000003:6001:42000121:77:…
…
```

Record type: State, event, communication

Who ➔ Application:process:thread

Event ➔ t_{event}:type:value

State➔ t_{start}:t_{end}:state

Comm Source ➔ $t_{logical\ send}$:$t_{physical\ send}$

Size and tag

Comm Dest ➔ $t_{logical\ recv}$: $t_{physical\ recv}$

CPU

Trace records (.prv)

- Not necessarily all types of records present in a trace
- Events:
  - Some predefined:
    - User functions (type 60000019). Value function identifier on entry. 0 on exit
    - MPI point to point calls (type 50000001). Value MPI call identifier on entry, 0 on exit
    - Hardware counters: PAPI cycles type 42000000, value is number of cycles since previous measurement
    - Flushing: type 40000003 value 1 on entry, 0 on exit
  - User defined:
- States: General type of activity:
  - Predefined: Idle loop, running, …..
- Communications:
  - Logical: user request to communicate
  - Physical start/end of actual data transfer

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Symbolic information (.pcf)

- Information of what is on the trace.
- Analysis can be performed even without it (in case of privacy concerns)

```
DEFAULT_OPTIONS

LEVEL                   THREAD
UNITS                   NANOSEC
LOOK_BACK               100
SPEED                   1
FLAG_ICONS              ENABLED
NUM_OF_STATE_COLORS     400
YMAX_SCALE              18
```

```
DEFAULT_SEMANTIC

THREAD_FUNC             State As Is

STATES
0    Idle
1    Running
2    Not created
3    Waiting a message
4    Blocked
5    Thd. Synchr.
6    Wait/WaitAll
7    Sched. and Fork/Join
8    Test/Probe
9    Blocking Send
10    Immediate Send
11    Immediate Receive
...
```

```
EVENT_TYPE
0   50000001    MPI Point-to-point
VALUES
3   MPI_Isend
4   MPI_Irecv
5   MPI_Wait
6   MPI_Waitall
0   End
```

```
EVENT_TYPE
0   50000002    MPI Collective Comm
VALUES
7   MPI_Bcast
8   MPI_Barrier
0   End


EVENT_TYPE
0   50000003    MPI Other
VALUES
19   MPI_Comm_rank
20   MPI_Comm_size
22   MPI_Comm_dup
23   MPI_Comm_split
31   MPI_Init
0   End
```

Trace control xml

mpitrace.xml

```
<?xml version='1.0'?>

<trace enabled="yes" home="/gpfs/apps/CEPBATOOLS/64.hwc">

  <mpi enabled="yes">
    <counters enabled="yes" />
  </mpi>

  <openmp enabled="no">
    <locks enabled="no" />
    <counters enabled="yes" />
  </openmp>

  <callers enabled="yes">
    <mpi enabled="yes">1-3</mpi>
  </callers>

  <user-functions enabled="no" list="/home/bsc41/bsc41273/user-functions.dat">
    <max-depth enabled="no">3</max-depth>
    <counters enabled="yes" />
  </user-functions>

...
```

Activate/not MPI/OpenMP tracing and features

Add call stack info (number of levels) at tracing points

Add instrumentation at specified user functions
Requires Dyninst based mpitrace

Trace control xml

mpitrace.xml (cont)

```
…

  <counters enabled="yes">
    <cpu enabled="yes" starting-set-distribution="1">
      <set enabled="yes" domain="all" changeat-globalops="5">
        PAPI_TOT_INS,PAPI_TOT_CYC,PAPI_L1_DCM
      </set>
      <set enabled="yes" domain="user" changeat-globalops="5">
        PAPI_TOT_INS,PAPI_FP_INS,PAPI_TOT_CYC
      </set>
    </cpu>

    <network enabled="yes" />

    <resource-usage enabled="yes" />
  </counters>

…
```

Specification of counters emitted to trace

When to rotate between groups

Groups

Interconnection network counters
Just at end of trace because of large acquisition overhead

OS info (context switches,….)

Trace control xml

mpitrace.xml (cont)

```
…

  <storage enabled="no">
    <trace-prefix enabled="yes">TRACE</trace-prefix>
    <size enabled="no">5</size>
    <temporal-directory enabled="yes" make-dir="no">/scratch</temporal-directory>
    <final-directory enabled="yes" make-
dir="no">/gpfs/scratch/bsc41/bsc41273</final-directory>
    <gather-mpits enabled="no" />
  </storage>

  <buffer enabled="yes">
    <size enabled="yes">150000</size>
    <circular enabled="no" />
  </buffer>

  <trace-control enabled="yes">
    <file enabled="no" frequency="5m">/gpfs/scratch/bsc41/bsc41273/control</file>
    <global-ops enabled="no"></global-ops>
  </trace-control>

…
```

Control of emitted trace …

… name, tmp and final dir …

… max (MB) per process size (stop tracing when reached)

Size of in core buffer (#events)

External activation of tracing (creation of file will start tracing)

Trace control xml

mpitrace.xml (cont)

```
…

  <others enabled="yes">
    <minimum-time enabled="no">10m</minimum-time>
    <terminate-on-signal enabled="no">USR2</terminate-on-signal>
  </others>

  <bursts enabled="no">
    <threshold enabled="yes">500u</threshold>
    <counters enabled="yes" />
    <mpi-statistics enabled="yes" />
  </bursts>

  <cell enabled="no">
    <spu-file-size enabled="yes">5</spu-file-size>
    <spu-buffer-size enabled="yes">64</spu-buffer-size>
    <spu-dma-channel enabled="yes">2</spu-dma-channel>
  </cell>

</trace>
```

Stop tracing after elapsed time …

… or when sighnal received

… emit only computation bursts of a minimal duration …

… plus summarized MPI events

Cell specifics

Extrae Environment variables

EXTRAE_ON / EXTRAE_CONFIG_FILE Enables instrumentation

EXTRAE_HOME Points where the Extrae is installed.

EXTRAE_PROGRAM_NAME Specify the prefix of the resulting intermediate trace files

EXTRAE_TRACE_TYPE Choose whether the tracefile is intended for Paraver or Dimemas

EXTRAE_DISABLE_MPI, EXTRAE_DISABLE_OMP, EXTRAE_DISABLE_PACX Disable MPI | OpenMP | PACX instrumentation

EXTRAE_COUNTERS Just one set can be defined. Counters separated by commas.

EXTRAE_FUNCTIONS_COUNTERS_ON, EXTRAE_MPI_COUNTERS_ON, EXTRAE_OMP_COUNTERS_ON, EXTRAE_PACX_COUNTERS_ON Specify if the performance counters should be collected when a function | MPI | OpenMP | PACX event is emitted

EXTRAE_BUFFER_SIZE Set the number of records that the instrumentation buffer can hold before flushing them

EXTRAE_FILE_SIZE Set the maximum size (in Mbytes) for the intermediate trace file

EXTRAE_MINIMUM_TIME Specify the minimum amount of instrumentation time

EXTRAE_CIRCULAR_BUFFER

EXTRAE_CONTROL_FILE, EXTRAE_CONTROL_GLOPS, EXTRAE_CONTROL_TIME

EXTRAE_DIR Specifies where temporal files will be created during instrumentation

EXTRAE_FINAL_DIR Specifies where files will be stored when the application ends

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Extrae Environment variables

EXTRAE_INITIAL_MODE Choose whether the instrumentation runs in in detail or
     in bursts mode

EXTRAE_BURST_THRESHOLD Specify the threshold time to filter running bursts

EXTRAE_MPI_STATISTICS, EXTRAE_PACX_STATISTICS Set to 1 if basic MPI | PACX statistics must be collected in burst mode

EXTRAE_MPI_CALLER, EXTRAE_PACX_CALLER select levels of the call stack to dump

EXTRAE FUNCTIONS
EXTRAE_OMP_LOCKS Set to 1 if locks have to be instrumented

EXTRAE_NETWORK_COUNTERS Set to 1 to dump network performance counters at flush points
     (Only available in systems with Myrinet GM/MX networks)
EXTRAE_RUSAGE Instrumentation emits resource usage at flush points if set to 1

EXTRAE_SPU_DMA_CHANNEL Choose the SPU-PPU dma communication channel
EXTRAE_SPU_BUFFER_SIZE Set the buffer size of the SPU side
EXTRAE_SPU_FILE_SIZE Set the maximum size for the SPU side (default: 5Mbytes)

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Merging individual files

```
Usage: ./mpi2prv inputfile1 ... [--] inputfileN [ -o <OutputFile>]
       ./mpi2prv -f TRACE.mpits [ -o <OutputFile>]
Options:
    -h         Get this help.
    -v         Increase verbosity.
    -o file    Output trace file name.
    -e file    Uses the executable file to obtain some information.
    -f file    MpitFILE File with the names of the ".mpit" input files.
    -syn       Synchronize traces at the end of MPI_Init.
    -syn-node Synchronize traces using node information.
    -no-syn    Do not synchronize traces at the end of MPI_Init.
    -maxmem M Uses up to M megabytes of memory at the last step of merging process.
    -dimemas  Force the generation of a Dimemas trace.
    -paraver  Force the generation of a Paraver trace.
    -s file   Indicates the symbol file attached to the *.mpit files.
    -d         Sequentially dumps the contents of every *.mpit file.
    -extended-glop-info
              Each global operation adds additional information records.
    -split-states
              Do not merge consecutives states that are the same.
    -skip-sendrecv
              Do not emit communication for SendReceive operations.
    -[no-]unique-caller-id
              Choose whether use a unique value identifier for different callers.
    --        Take the next trace files as a diferent parallel task.
```


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Adding references to sources (steps)

- Build the binary with -g option
- Specify to add callers on the XML configuration file

```
<callers enabled="yes">
  <mpi enabled="yes">1-3</mpi>
  <pacx enabled="no">1-3</pacx>
  <sampling enabled="no">1-5</sampling>
</callers>
```

- Merge the mpit files including the -e option
  - `${EXTRAE_HOME}/bin/mpi2prv -f TRACE.mpits -e <path2binary> -o output.prv`
- Use the available configuration files

Examples of MN set-up available at
/gpfs/apps/CEPBATOOLS/tracing-setup/WithCallers

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Extrae API

- Emit event records to the trace
  - functions
    - `void Extrae_event (int event, int value)`
    - `void Extrae_nevent (unsigned int count, unsigned int *types, unsigned int *values)`
    - `void Extrae_counters (void)`
    - `void Extrae eventandcounters (int event, int value)`
    - `void Extrae_neventandcounters (int event, int value)`
    - `void Extrae_network_counters (void)`
    - `void Extrae_network_routes (int task)`

ONLY @ MareNostrum

- useful to
  - set reference points for comparison between traces
  - visualize variable value evolution

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Extrae API

- Control calls
  - functions
    - `void Extrae_shutdown (void)`
    - `void Extrae_restart (void)`
    - `void Extrae_init (void)`
    - `void Extrae_fini (void)`
  - useful to
    - reduce size of the tracefile
    - concentrate on the interesting parts
  - function
    - `void Extrae_previous_hwc_set (void)`
    - `void Extrae_next_hwc_set (void)`
  - useful to
    - Change the active counter set (defined in XML)
  - function
    - `void Extrae_set_tracing_tasks (int from, int to)`
    - `void Extrae_set_options (int options)`
  - useful to
    - Control what/who it is instrumented

Automatically inserted in MPI applications

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Instrumentation hands-on

- Obtain a tracefile for a small problem size
  - Follow steps to include source references
  - Addapt the Extrae configuration file
  - Instrument a run of your application
  - Merge the mpits with mpi2prv

- Repeat the hands-on guidelines of the Paraver session with your application trace

- Obtain a new trace doubling the number of tasks and compare both executions. Do the application achieve a good scalability? Do you identify any problem?
  - Do all the phases of the program achieve the same scalability?
  - Compare the useful duration histogram
  - Compare the message size histogram