



ORNL is managed by UT-Battelle, LLC for the US Department of Energy



Extrae/Paraver

- Developed by Barcelona Supercomputing Center
- Extrae for instrumentation
- Paraver for visualization and performance analysis
- Installed version on Summit: v3.7.1
- Module: extrae
- Web site: <u>https://tools.bsc.es/extrae</u> <u>https://tools.bsc.es/paraver</u>



Capability Matrix - Extrae

Capability	Profiling	Tracing	Notes/Limitations
MPI, MPI-IO	Yes	Yes	
OpenMP CPU	Yes	Yes	Only GNU
OpenMP GPU	Yes	Yes	Only with GNU compiler, no OpenACC
OpenACC	No	No	
CUDA	Yes	Yes	Not advanced
POSIX I/O	ŚŚ	ŚŚ	
POSIX threads	Yes	Yes	
Memory – app-level	Yes	Yes	Need to use dynamic allocation
Memory – func-level	Yes	Yes	Need to use dynamic allocation
Hotspot Detection	Yes	Yes	
Variance Detection	Yes	Yes	
Hardware Counters	Yes	Yes	



Compilation

- Extrae is a bit more complicate to start using it compared to many other tools
- We can have dynamic or static compilation
 - For static, it is required to recompile
 - For dynamic is required to compile with -g, it works even without -g but less information will be instrumented:



How does Extrae work?

- Symbol substitution through LD_PRELOAD
 - We need to use specific libraries based on programming language/model
- Dynamic instrumentation (based on DynInst)

• Static link



Trace Generation Workflow





Library Selection

- Choose a library depending on the application type
 - The suffix "f" is for Fortran codes

Library	Serial	MPI	OpenMP	pthread	CUDA
libseqtrace	\checkmark				
libmpitrace[f] ¹		\checkmark			
libomptrace			\checkmark		
libpttrace				\checkmark	
libcudatrace					\checkmark
libompitrace[f] ¹		\checkmark	\checkmark		
libptmpitrace[f] ¹		\checkmark		\checkmark	
libcudampitrace[f] 1		\checkmark			\checkmark

¹ include suffix "f" for Fortran codes





2/3D tables (Statistics)

Trace visualization/analysis

+ trace manipulation

Timelines

Goal = Flexibility No semantics Programmable

Comparative analyses Multiple traces Synchronize scales



Extrae XML configuration - MPI

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

```
<openmp enabled="yes" ompt="no">
   <locks enabled="no" />
        <taskloop enabled="no" />
        <counters enabled="yes" />
   </openmp>
```

```
<cuda enabled="no" />
```

```
<pthread enabled="no">
<locks enabled="no" />
<counters enabled="yes" />
</pthread>
```

<callers enabled="yes"> <mpi enabled="yes">1-3</mpi> <sampling enabled="no">1-5</sampling> <dynamic-memory enabled="no">1-3</dynamic-memory> <input-output enabled="no">1-3</dynamic-memory> <isyscall enabled="no">1-3</syscall>

</callers>

<counters enabled="yes"> <cpu enabled="yes" starting-set-distribution="1"> <set enabled="yes" domain="all" changeat-time="0"> PAPI_TOT_INS, PAPI_TOT_CYC, PAPI_FP_OPS </set> <set enabled="no" domain="all" changeat-time="0"> PAPI_TOT_INS, PAPI_TOT_CYC, PAPI_SR_INS, PAPI_FP_INS <sampling enabled="no" period="1000000000">PAPI_TOT_CYC</sampling> </set> </cpu> <network enabled="no" /> <memory-usage enabled="no" /> </counters>

<buffer enabled="yes">
<size enabled="yes">5000000</size>
<circular enabled="no" /></buffer>

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

<sampling enabled="no" type="default" period="50m" variability="10m" />

<dynamic-memory enabled="no"> <alloc enabled="yes" threshold="32768" /> <free enabled="yes" /> </dynamic-memory>



Execution and Merging

- jsrun -n 64 -r 8 -a 1 -c 1 ./trace.sh ./miniWeather_mpi
- trace.sh:

```
#!/bin/bash
export EXTRAE_HOME=/sw/summit/extrae/3.7.1/rhel7.5_gnu6.4.0
export EXTRAE_CONFIG_FILE=/full_path/extrae.xml
export LD_PRELOAD=${EXTRAE_HOME}/lib/libmpitrace.so:$LD_PRELOAD
$*
```

• jsrun -n 64 -r 8 -a 1 -c 1 mpimpi2prv -f TRACE.mpits -e miniWeather_mpi



After the execution with merging

- A folder set-X where X is number 0,1, etc. with the traces, one folder for every 256 MPI processes
- Files based on the merging output, *.prv, *.pcf, *.row, the first one is the merged trace and the rest information about the trace and the events.
- Now you need to visualize the trace for performance analysis.
- We use the tool Paraver, it is available for Linux, Mac, Windows and already pre-compiled (<u>https://tools.bsc.es/downloads</u>), quite difficult to be built on Power processor. Available on Rhea or your computer.



Paraver on Rhea

% ssh –Y <u>username@rhea.ccs.ornl.gov</u> % module load paraver % wxparaver

Location for configuration files: /sw/rhea/paraver/cfgs/

Is /sw/rhea/paraver/cfgs/

burst_mode clustering counters_PAPI CUDA folding General Java mpi OmpSs OpenCL OpenMP pthread sampling+folding scripts software_countersspectral uninstall.sh



Paraver – Load trace

•	0	X Paraver
File	Hints Help	
	👒 🔳 🗙 🎸 🎲	
Worksp	aces	
Non	9	
Windo	v browser	
All Tr	aces	•

0 0	
le Hints Help	
Load <u>T</u> race	Ctrl+0
Previous Traces	•
Unload Traces	
Load Configuration	
Previous Configurations	►
Save Configuration	
Load Session	Ctrl+L
Save Session	Ctrl+S
Preferences	
Quit	Ctrl+Q

Load trace

iles & Window Properties	
🕨 🛅 run	
🕨 🛅 sbin	
🛅 srv	
🕨 🛅 sw	
🕨 🚞 sys	
🕨 🛅 tmp	
🕨 🛅 usr	
🕨 🛅 var	
Paraver files	•
Automatic Redraw	Force Redraw

Files & Window Properties	
▶ □ run ▶ □ sbin	
srv • 💽 sw • 🗖 svs	
 imp imp imp 	
Var	
Paraver files	Force Redraw



Paraver - Filter trace

Name	▼ Size	Modified			
🛅 old		12:31			
set-0		12:01			
📄 miniWeather_mpi.prv	3.3 GB	12:09		000	🔀 Reduce trace size
			Reduce trace size	?	The size (3267 MB) of the trace /gpfs/alpine/gen110/scratch/gmarkoma/extrae/miniweather/c/ miniWeather_mpi.prv
			\longrightarrow		exceeds the maximum loadable defined in Preferences. Would you like to cut/filter the trace?
					No Cancel Yes
	Paraver trace (*.prv	r;*.prv.gz) ▼			Click Yes
	Cancel	Open			



Paraver - Filter trace

Click Browse and load the filter.xml file

	🔀 Cut & Filter			
Traces				
Input	miniWeather_mpi.prv	Brows	e	
Output	miniWeather_mpi.filter1.prv	Brows	e	
	 Load the processed trace Run application with the processed trace 		/	
Cut/Filter Parameters —				
Configuration file	filter.xml	Brow	se	
	Execution chain 1. Filter 2 Cutter 3 Software Counters	1	Save	
Cutter Filter Softwa	are Counters			
Discard Records				Cutter Filter Software Counters
🗌 State 🕑 Event 🗌	Communication			/ Discard Records
Keep states				State 🖌 Event 🗌 Communication
Idle				Keen states
🛃 Running		Select all		
Not created		Unselect all		Inmediate Send Select all
Waiting a message	e			Inmediate Receive
Bloking Send		Min. burst time 50000		
Ind. Synchr.				Tracing Disabled Min. burst time 50000
Events				Others
		Add		Fuente
		Add		
		Delete		
		Discard		
Keep communications				
Minimum size	1		A Bytes	
Minimum size	1		▼ ^{bytes}	
		Cancel	Apply	



Paraver – Visualize trace

Click Browse and load the filter.xml file







Paraver – Investigating trace



Remove the communication links





Open slide master to edit

Paraver - Zoom



Zoom, left click with mouse and select area moving the cursor horizontally towards right and decide which part we want to study





Paraver – Computation configuration file

• We load h_comp_time.cfg, File -> Load configuration

File	Hints	Help	
Loa	ad Trace		•
Pre	evious Tr	aces	
Un	load Trac	ces	
Loa	ad Config	guration	•
Pre	evious Co	onfigurations	
Sav	ve Config	guration	
Loa	ad Sessio	on	жı
Sav	ve Sessio	on	ж



JDON'T UNDERSTAND

🔲 Useful Duration @ miniWeather_mpi.filter1.prv HREAD 1.1. HREAD 1.5. THREAD 1.9. THREAD 1.13 HREAD 1.17 HREAD 1.21 THREAD 1.25 THREAD 1.29 HREAD 1.33 HREAD 1.37 HREAD 1.41 HREAD 1.45 HREAD 1.49 HREAD 1.53 HREAD 1.57 THREAD 1.61.1 THREAD 1.64.1

CAK RIDGE

19

Click on Open Control Window

Paraver – Computation configuration file

• We load h_comp_time.cfg, File -> Load configuration







We can see some iterations

😕 🗉 Usefi	ul Duration @ miniWeat	her_mpi.filter1.prv:		
THREAD 1.1.1				
THREAD 1.5.1				
THREAD 1.9.1				
THREAD 1.13.1				
THREAD 1.17.1				
THREAD 1.21.1				
THREAD 1.25.1				
THREAD 1.29.1				
THREAD 1.33.1				
THREAD 1.37.1				
THREAD 1.41.1				
THREAD 1.45.1				
THREAD 1.49.1				
THREAD 1.53.1				
THREAD 1.57.1				
THREAD 1.61.1 THREAD 1.64.1				
	14.663.197 us			15.716.223 us



Paraver – Extract part of the original trace

Select Filter Trace •

Useful Duration		
Useful Duration		

😣 🗉 🛛 Cut & Filte	er		
Traces			
Input	miniWeather_mpi.prv		Browse
Output	miniWeather, mpi chop	DEV	Browse
Odtput	Vedener_mpr.cnopg.		Drowse
	Run application with th	ace	
		le processed trace	
Cut/Filter Parame	ters		
Configuration file			Browse
	Execution chain		
	1 Cutter		Save
	3 - Software Count	ers	
			I
Cutter Filter	Software Counters		
Trace Limits			
Out by time	e Begin		
O Cut by time	e % End		
	End		
Tasks			
Select Re	egion All v	Vindow	All Trace
Trace Options			
🗌 Use original	ltime	Remove first sta	ate
🗌 🗌 Don't break	states	🗌 Remove last sta	te
		(Cancel Apply
	Paraver		
Window b			
Window b	lowser	a da (asia itu ta abbaa	
/nome/	gmarkomanolis/Downic	ads/miniweather_	mpi.cnop1.prv

- Select for Input the original trace Select cut for the execution chain ٠
- ٠
- Trace options: Use original time to be able to compare between traces and remove last
- state Click Select region and mark the area to cut from the original trace Click Apply, the trace will be created and loaded ٠
- ٠

Cutter Filter Software Counters		
Select Region All Window	All Trace	
Trace Options		
☑ Use original time	irst state	
🔲 Don't break states 🛛 🐼 Remove	ast state	
Keep boundary events		
Output Taraa		
Maximum trace size	A MP	
0	MD	
	Capital	
	Cancet Apply	
😠 🗊 Useful Duration @ miniWea	ther mpi.filter1.prv	
THREAD 1.1.1		
THREAD 1.5.1		
THREAD 1.9.1		
THREAD 1.13.1		
THREAD 1.17.1		
THREAD 1.25.1		
THREAD 1.29.1		
THREAD 1.33.1		
THREAD 1.37.1		
THREAD 1.41.1		
THREAD 1.45.1		
THREAD 1.49.1		
THREAD 1.57.1		
THREAD 1.61.1		
THREAD 1.64.1 14,663,197 us		15,716,223



Paraver – MPI Profile

Select Hints -> MPI -> MPI Profile





C 🗋 3D 🤇	2 🔍 🔳 🖪	• 🖪 🔳 🖄	Σ%					
	Outside MPI	MPI_Isend	MPI_Irecv	MPI_Waitall	MPI_Bcast	MPI_Allreduce	MPI_Comm_rank	MPI_Comm_dup
THREAD 1.1.1	98.29 %	0.12 %	0.11 %	0.27 %	0.01 %	0.50 %	0.00 %	0.01 %
THREAD 1.2.1	98.45 %	0.12 %	0.11 %	0.13 %	0.01 %	1.06 %	0.00 %	0.01 %
THREAD 1.3.1	98.28 %	0.11 %	0.11 %	0.30 %	0.01 %	1.06 %	0.00 %	0.01 %
THREAD 1.4.1	98.28 %	0.11 %	0.11 %	0.30 %	0.01 %	1.06 %	0.00 %	0.01 %
THREAD 1.5.1	98.28 %	0.11 %	0.11 %	0.29 %	0.01 %	1.06 %	0.00 %	0.01 %
THREAD 1.6.1	98.45 %	0.11 %	0.11 %	0.13 %	0.01 %	1.06 %	0.00 %	0.01 %
THREAD 1.7.1	98.28 %	0.11 %	0.11 %	0.29 %	0.01 %	1.06 %	0.00 %	0.01 %
THREAD 1.8.1	98.28 %	0.12 %	0.11 %	0.29 %	0.01 %	1.06 %	0.00 %	0.01 %
THREAD 1.9.1	98.28 %	0.12 %	0.11 %	0.29 %	0.01 %	0.40 %	0.00 %	0.01 %
HREAD 1.10.1	98.45 %	0.11 %	0.11 %	0.13 %	0.01 %	0.58 %	0.00 %	0.01 %
HREAD 1.11.1	98.28 %	0.11 %	0.11 %	0.30 %	0.01 %	0.58 %	0.00 %	0.01 %
HREAD 1.12.1	98.28 %	0.11 %	0.11 %	0.30 %	0.01 %	0.58 %	0.00 %	0.01 %
HREAD 1.13.1	98.28 %	0.11 %	0.11 %	0.29 %	0.01 %	0.58 %	0.00 %	0.01 %
HREAD 1.14.1	98.44 %	0.11 %	0.11 %	0.13 %	0.01 %	0.58 %	0.00 %	0.01 %
HREAD 1.15.1	98.28 %	0.11 %	0.11 %	0.29 %	0.02 %	0.58 %	0.00 %	0.01 %
HREAD 1.16.1	98.28 %	0.11 %	0.11 %	0.29 %	0.02 %	0.58 %	0.00 %	0.01 %
HREAD 1.17.1	98.28 %	0.12 %	0.11 %	0.28 %	0.02 %	0.58 %	0.00 %	0.01 %
HREAD 1.18.1	98.44 %	0.11 %	0.11 %	0.13 %	0.02 %	0.58 %	0.00 %	0.01 %

Scroll down

Click Hints -> MPI profile -> Histogram Zoom

The average under the column Outside MPI represents the parallel efficiency, the value Avg/Max is the load balance and the Max is the communication efficiency

🐵 🗊 🛛 MPI call 🛛	profile @ miniV	Veather_mp	i.chop1.prv					
ie d 3d 🔾	. 🔍 🔳 н	H II 🛪	Σ ½					
IRREAU 1.33.1	50.20 /0	<u> </u>	0.11.20	0.25 /0	0.02 /0	0.37 /0	0.00 %	0.01 /0
THREAD 1.54.1	98.44 %	0.11 %	0.11 %	0.13 %	0.02 %	0.58 %	0.00 %	0.01 %
THREAD 1.55.1	98.28 %	0.11 %	0.11 %	0.29 %	0.02 %	0.58 %	0.00 %	0.01 %
THREAD 1.56.1	98.28 %	0.12 %	0.11 %	0.28 %	0.02 %	0.58 %	0.00 %	0.01 %
THREAD 1.57.1	98.29 %	0.12 %	0.11 %	0.28 %	0.02 %	0.58 %	0.00 %	0.01 %
THREAD 1.58.1	98.44 %	0.11 %	0.11 %	0.13 %	0.02 %	0.58 %	0.00 %	0.01 %
THREAD 1.59.1	98.28 %	0.11 %	0.11%	0.30 %	0.02 %	0.58 %	0.00 %	0.01 %
THREAD 1.60.1	98.28 %	0.11 %	0.11 %	0.29 %	0.02 %	0.58 %	0.00 %	0.01 %
THREAD 1.61.1	98.28 %	0.11 %	0.11 %	0.29 %	0.02 %	0.58 %	0.00 %	0.01 %
THREAD 1.62.1	98.44 %	0.11 %	0.11 %	0.13 %	0.01 %	0.58 %	0.00 %	0.01 %
THREAD 1.63.1	98.28 %	0.11 %	0.11 %	0.29 %	0.02 %	0.58 %	0.00 %	0.01 %
THREAD 1.64.1	98.28 %	0.12 %	0.11 %	0.29 %	0.01 %	0.58 %	0.00 %	0.01 %
Total	6.292.58 %	7.28 %	7.13 %	15.79 %	1.12 %	40.08 %	0.28 %	0.60 %
Average	98.32 %	0.11 %	0.11%	0.25 %	0.02 %	0.63 %	0.00 %	0.01 %
Maximum	98.45 %	0.12 %	0.12 %	0.30 %	0.02 %	1.06 %	0.00 %	0.01 %
Minimum	98.28 %	0.11 %	0.11%	0.12 %	0.01 %	0.40 %	0.00 %	0.01 %
StDev	0.07 %	0.00 %	0.00 %	0.07 %	0.00 %	0.16 %	0.00 %	0.00 %
Avg/Max	1.00	0.96	0.97	0.83	0.75	0.59	0.96	0.90

MPI_Isend



Paraver - Analyzing the trace - MPI Profile

🜔 D 30 🔍 🍔 📕 н Η 📶 💥 Σ ½



	THREAD 1.1.1	THREAD 1.2.1	THREAD 1.3.1	THREAD 1.4.1	THREAD 1.5.1	THREAD 1.6.1	THREAD 1.7.1	THREAD 1.8.1	THREAD
	THREAD 1.1.1	THREAD 1.2.1	THREAD 1.3.1	THREAD 1.4.1	THREAD 1.5.1	THREAD 1.6.1	THREAD 1.7.1	THREAD 1.8.1	THREAD
THREAD 1.1.1	-	12,800	-	-	-	-	-	-	
THREAD 1.2.1	12,800	-	12,800	-	-	-	-	-	
THREAD 1.3.1	-	12,800	-	12,800	-	-	-	-	
THREAD 1.4.1	-	-	12,800	-	12,800	-	-	-	
THREAD 1.5.1	-	-	-	12,800	-	12,800	-	-	
THREAD 1.6.1	-	-	-	-	12,800	-	12,800	-	
THREAD 1.7.1	-	-	-	-	-	12,800	-	12,800	
THREAD 1.8.1	-	-	-	-	-	-	12,800	-	1
THREAD 1.9.1	-	-	-	-	-	-	-	12,800	
THREAD 1.10.1	- L	-	-	-	-	-	-	-	1
THREAD 1.11.1	L -	-	-	-	-	-	-	-	
THREAD 1.12.1	L -	-	-	-	-	-	-	-	
THREAD 1.13.1	L -	-	-	-	-	-	-	-	
THREAD 1.14.1	L -	-	-	-	-	-	-	-	
THREAD 1.15.1	L -	-	-	-	-	-	-	-	
THREAD 1 16 1		_	_	_	-	_	-	_	

Select Window properties -> Communication for Type and Maximum bytes sent for Statistic.



Paraver - Computation

- Load the 2dh_usedulduration.cfg for a histogram of the duration for the computation regions A lot of areas are not constituted by vertical lines which shows load imbalance. We explore in the next slide what is inside the red circle We select Open Filtered Window and we zoom in the area of red circle





Paraver - Computation

- We zoom in the first area, we compare with the MPI calls and the 2dp_line_call.cfg Only the processes 2-8 execute this part and seems that is not instrumented, thus, it could be from an external library

😣 🗊 U	Jseful D	uration 2DZoom range [3	34.35,445.1	71) @ mini\	Weather_m	pi.chop1.p	rv	
THREAD 1.2	2.1				I			
THREAD 1.3	3.1							
THREAD 1.4	4.1							
THREAD 1.5	5.1							
THREAD 1.6	6.1							
THREAD 1.5	7.1							
THREAD 1.8	8.1							
	14,24	8,286 us		14,613,791	us - 14,613	,791 us		16,364,368 u





Paraver Useful Instructions

Load the 2dh_useful_instructions.cfg

🛞 🗇 2dh useful i structions @ miniWeather_mpi.chop1.prv	
	-
	-
	-
	-

[2,355,420.55..2,391,631.99)





Paraver – Extract part from the original trace



nput	miniWeather_mpi	i por		
nput	miniWeather_mpi	I DEV		
Sutmut		i.prv		Browse
Julpul	miniWeather_mpi	i.chop2.prv		Browse
	Load the proces	ssed trace		
0	Run application	with the processed trace		
ut/Filter Parameters				
onfiguration file				Browse
	Execution ch	ain		
	🛃 1 Cutter			
	2 Filter		Û	Save
	🗌 3 Softwa	are Counters		
Cutter Filter Softw	ware Counters			
Trace Limits				
 Cut by time 	Begi	in		
O Cut by time %	End			
Tasks				
Select Regi	ion	All Window	All Trac	ce
Trace Options				
Use original time	e	Remove first	state	
Don't break stat	es	Remove last	state	
Keep boundary	events			
Output Trace				
Maximum trace size	0			▲ MB

After we click "Select Region..." then select the area from the already opened filtered trace



X New window #4 @ miniWeather mpi.filter1.prv





Paraver – Profile per calling line

We load the 2dp_line_call.cfg, we select open control window and synchronize the new window





Paraver – Late receivers

We load the late_receivers.cfg and the 2dp_line_call.cfg







D IN STR er MPI caller line @ minWeather_mpi.chop2.prv Copy Copy VC Default Str Str
Str

CAK RIDGE

29

Open slide master to edit

Paraver – Late senders

We load the receiver_from_late_sender.cfg and the 2dp_line_call.cfg





MiniWeather MPI+OpenMP

- jsrun -n 64 -r 8 -a 1 -c 2 ./trace_openmp.sh ./miniWeather_mpi_openmp
- Trace_opnemp.sh: #!/bin/bash export EXTRAE_HOME=/sw/summit/extrae/3.7.1/rhel7.5_gnu6.4.0 export EXTRAE_CONFIG_FILE=/gpfs/alpine/.../c/extrae_openmp.xml export LD_PRELOAD=\${EXTRAE_HOME}/lib/libompitrace.so:\$LD_PRELOAD

Run the desired program

\$*

• jsrun -n 64 -r 8 -a 1 -c 2 mpimpi2prv -f TRACE.mpits -e ./miniWeather_mpi_openmp



Parallel Loops

- Create a new chop file as described before
- Load the parallel_loops.cfg and zoom





Parallel Loops

- Create a new chop file as described before
- Load the parallel_loops.cfg and zoom





Load Balance

- Load OpenMP/analysis/load_balance.cfg
- Load the parallel_loops.cfg and zoom





TASK 1.18 set_halo_values_x = 0.93



Load Balance

Load OpenMP/analysis/load_balance.cfg





Load Balance

Load OpenMP/views/parallel_functions_useful.cfg and zoom





Paraver - Flush





CAK RIDGE National Laboratory

Paraver – Chop bigger area

Activate communication lines Right Click -> View -> Communication lines

