

Barcolona Supercomputing Center Centro Nacional de Supercomputación

# Obtaining Extremely Detailed Information at Scale

Jesus Labarta Barcelona Supercomputing Center

### **Topics**

- Instrumentation + Sampling
  - Objective
    - Can we get very fine grain Information ...
    - ... of time distribution of metrics ...
    - ... with little overhead?
  - Work by Harald Servat

- Use of MRNET
  - Miminize the captured data/information ratio
    - Statistics, profile, traces
  - Work by German Llort





### Instrumentation

- Events correlated to specific program activity
  - Start/exit iterations, functions, loops,...



- Different intervals:
  - May be very large, may be very short
  - Variable precision
- Captured data:
  - Hardware counters, call arguments, call path,....





### Sampling



• Time (or counter) overflow



- Controlled granularity:
  - Sufficiently large to minimize overhead
  - Guaranteed acquisition interval/precision
- Statistical projection
  - %Counts = %time (or metric)
  - Assuming no correlation, sufficiently large #samples





### Instrumentation + sampling



Both



- Guaranteed interval
- Captured data:
  - Hardware counters (since previous probe)
  - call path
  - Call arguments in some probes
- How to use it?





### **New roles**



- Instrumentation  $\rightarrow$  Reference
  - Identify different instances of a region for which to obtain detailed time evolution of metrics
    - Stationary behaviour assumed
  - Target region:
    - Iteration
    - Routine
    - Routine excluding MPI calls
    - . . .







### **New roles**

- Sampling role  $\rightarrow$  relative data
  - Guarantee granularity
  - Provide data to increase granularity







### **Folding counters: Projecting**

- Cumulative count since reference
- Internal variance
- Variance in duration
  - Eliminate outliers
  - Scale







### **Folding counters: Fitting**

- Eliminate outliers
- Kriging interpolation









### **Timeline display**

- Present
  - Analytical expression
  - Plots, statistics
    - Time, IPC,...
  - Timelines
- Performance counters:
  - Sample again fitted function and inject synthetic events into trace
- Call stack
  - Truncated by specifying routines of interest



THREAD 1,11,1 - AADA	
ואאראל אין	
	Ī
A LANNON I LAND LITY & Good I TI MI has	
A 476419240 ns 519728001 ns 563036762 ns 606345523 ns 649654284 ns	
	_
Folded caller outside MPI @ bt.B.16.50mcycles.5b.sorted.prv 💿 🖉	8

Folded instructions nocolor @ bt.B.16.50mcycles.5b.sorted.prv

MMT

Folded caller	outside MPI@	bt.B.16.50mcyc	les.ob.sorted.p	rv 🛞 🙆	8
THREAD 1.1.1					
THREAD 1.2.1 -					
THREAD 1.3.1					
THREAD 1.4.1					
THREAD 1,5,1					
THREAD 1.6.1					
🗰 User function	x thread @ bt.	3.16.50mcvcles	.5b.sorted.prv		×
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 1.9.1					
THREAD 1,10,1					
THREAD 1.11.1					
THREAD 1.12.1					
INREHD 1,15,1 -					
THREAD 1,16,1 -					
473805,72 us	516864,63 us	559923,53 us	602982,43 us	646041.34 us	689100,24
and and a second se					





### Lots of detail form a single run with low overhead







11

### **Beware of correlations**

- 50 Mcycles seemed a reasonable interval
- Ends up being slightly correlated to application periodicity 😕
  - Can bias a pure statistical profile
  - Folding has potential to reduce this effect

98	mpres_z			e nue.				0 0 0
THREAD 1.1.1 -	P	P P	e e	P P			P P	PPP
	P	P P	P P	P P	P P	e e	P P	PPP
	P P	P P	P P	P		P, P, F	• P P	PPP
	P	P P	P P	P P	P P	e e	P P	PPP
THREAD 1.5.1 -	e e	P P	E E	P. F	P 1	e e e	e e	e e e
	P P			PP		PP	PP	
	PP		P P	P F				P P P
THREAD 1.8.1 -	PP			P		P P P	P P	PPP
THREAD 1,9,1 -	P P	PP		P P		PP	P P	P
THREAD 1.10.1 -	P	P P	P P	P P			P P	PPP
THREAD 1+11+1 -	P. P.	P P	P 1			PP	PP	
THREAD 1,12,1 -	P	e e	e e	P P	P P	P P	P P	PPP
THREAD 1.13.1 -	P	P P	PP	P B	P P	PP	PP	
THREAD 1.14.1 -	P	P P	P P	P P	P P	P P	PP	PPP
THREAD 1,15,1 -	P	PP	PP	PR	PP	PP	P P	
TUD500 4 44 4	P	PP	P P	P P	PP	-		
INKEHD 1+10+1 -								
126	83811732 ns	127852	75447 ns	128867391	51 ns 1	2988202876 ns	13089666	5591 ns
126	83811732 ns	127852	75447 ns	128867391	51 ns 1	2988202876 ns	13089666	5591 ns
126 126 126	83811732 ns	127852 thread (	75447 ns	128867391 6.50mcyc	les.5b.s	2988202876 ns	13089666	5591 ns
126 126 Userfu	83811732 ns	127852 thread (	25447 ns	128867391 6.50mcyc	51 ns 1 les.5b.sc	2988202876 n#	13089666	5591 ns
User fu THREAD 1,1,1	83811732 ns	127852 ( thread (	75447 ns	128867391 6.50 m c y c	ins 1	2988202876 n#	13089666	5591 ns
THREAD 1,10,1 - 126 User fu THREAD 1,1,1 THREAD 1,2,1 THREAD 1,3,1 THREAD 1,4,1	83811732 ns	127852 ( thread (	25447 ns	128867391 6.50mcyc	ins 1	2988202876 ns	13089666	5591 n <i>s</i>
THREAD 1,10,1     126     User fu     THREAD 1,1,1     THREAD 1,2,1     THREAD 1,2,1     THREAD 1,3,1     THREAD 1,5,1	83811732 ns	127852	25447 ns	128867391	iin⊴ i les.5b.so	2988202876 ns	13089666	5591 n <i>s</i>
Ize     Impact of the second se	nction x	: 127852	75447 ns	128867391	sins i les.5b.so	2988202876 ng	13089666	5591 ns
Implement   Item	nction x	: 127852 : thread (	75447 ns	128867391	ins 1	2968202876 ns	13089660	5591 n.s
IMPERIO 1.1.61     126     User fu     IHERAO 1.1.1     IHERAO 1.3.1     IHERAO 1.3.1     IHERAO 1.4.1     IHERAO 1.5.1     IHERAO 1.6.1     IHERAO 1.6.1     IHERAO 1.6.1	nction x	: 127852 : thread (	25447 ns	128867391	ins 1	2998202876 n;	13089666	5591 ns
THREAD 1.161     126     W User fu     THREAD 1.11     THREAD 1.2.1     THREAD 1.3.1     THREAD 1.4.1	83811732 ng	: 127852 : thread (	25447 ns	128867391	sins 1	2998202876 ns	13089666	591 ns
IMPERIO 1.151     126     W User fu     IHERO 1.1.1     IHERO 1.2.1     IHERO 1.3.1     IHERO 1.4.1     IHERO 1.5.1     IHERO 1.6.1     IHERO 1.6.1     IHERO 1.6.1     IHERO 1.6.1     IHERO 1.6.1     IHERO 1.6.1     IHERO 1.9.1     IHERO 1.9.1	83811732 n/	: 127852 ( thread @	25447 ns	128867791	sins 1	2996202876 ns	13089666	5591 n.s
Тинено 1.151 Тинево 1.111 Тинево 1.211 Тинево 1.311 Тинево 1.411 Тинево 1.411	nction x	: 127852	25447 ns	128867391	ilns 1	2988202876 ns	13089666	5591 ns () () () () () () () () () () () () () (
Immendi 1.15.1     126     Userfu     THREAD 1.2.1     THREAD 1.3.1     THREAD 1.3.1     THREAD 1.4.1	nction x	thread (	25447 n.s	128867391	itn: t	2985202876 ns	13089666	5591 ns
IMPERIO 1.151     126     W User fu     IHREAD 1.11     IHREAD 1.2.1     IHREAD 1.2.1     IHREAD 1.4.1     IHREAD 1.5.1     IHREAD 1.6.1     IHREAD 1.6.1     IHREAD 1.7.1     IHREAD 1.7.1     IHREAD 1.9.1     IHREAD 1.1.1.1     IHREAD 1.2.1     IHREAD 1.3.1     IHREAD 1.3.1	nction x	127652 ( thread @	25447 ns	128867391	šins i les.5b.sc	2988202876 ns	1308966	5591 n.s
THREAD 1.15.1     THREAD 1.1.1     THREAD 1.2.1     THREAD 1.3.1     THREAD 1.4.1	nction x	: 127852 : thread (	28447 ns	128667391	sins i les.5b.sc	2988208876 ns	1308966	591 ns
IMPERIO 1.15.1     126     Userfu     THREAD 1.2.1     THREAD 1.3.1     THREAD 1.4.1	nction x	thread (	25447 n:	128667391	ins i	2938202876 ns	13089666	5591 ns
IMPERIO 1.15.1     IMPERIO 1.11.1     IMPERIO 1.3.1     IMPERIO 1.4.1     IMPERIO 1.4.1     IMPERIO 1.5.1     IMPERIO 1.6.1     IMPERIO 1.15.1     IMPERIO 1.15.1     IMPERIO 1.16.1	53511732 ns	127852	25447 ns	128867391	ins i les.5b.sr	2986202876 ns	13089666	5591 n.s







## Maximizing information/data ratio





### **Distributed trace control**

- MRNet based mechanism
  - Local instrumentation on a circular buffer
  - Periodic MRNet front-end initiation of collection process
    - "Intelligent" determination of interval
  - Local algorithm
  - Reduction on tree, selection at root and propagation
    - Cluster a subset of all data
    - Classification of all data
    - Generation of global statistics
  - Locally emit trace events

Clustering tool





### **Clustering vs. classification**

- Clustering time drastically grows with number of points
- Selection of a subset of data to clusterize
  - Space: Select a few processes. Full time sequence
  - Random sampling: wide covering
- Remaining data: "nearest" neighbor classification



### **GROMACS 64 processes**



#### 64 processes



#### 25% random records



#### 15% random records

IUDEOB 1.4.1				_		_			_	
										<b></b>
11° 111 X 11 1							<b>-</b>			
			- <b>-</b> -	_						- <b>-</b> -
					- 25:					
				· <u></u> ·		·	<u></u>			- 12,000
			<b>25</b> .0	- 22.3			<b>99 1 3</b>			
	_	<u></u>	- <b>1</b>	i <b>sa</b> ing i	- Second -	e <b>en e</b> en	<b></b>	and a second		
						1				
			-			· · · · ·				
						· <u>- · ·</u> · ·				
			_	• ==					-	
			-							
	86493832	2.56	866	00482.17	5 11 4	86707131.	70.03	86813781.2	7	86920430.84 0
-										

#### 10% random records

183656717,29 us

84169535	.89 us	8427703	52.05 us	843	84528.20	us	8449202	+,36 us	845995	20.52 us

#### 8 random processes + 15% random records

183328173,95 us

Good quality Fast analysis

Jesus Labarta, Workshop on Tools for Petascale Computing, Lake Tahoe, July 2009



Ξ

183547202.85 us

183437688.40 us



### **Tracking evolution**

- Compare 2 clusterings cluster per cluster
  - Inscribe clusters into a rectangle with a 5% margin of variance. Match those that overlap.
  - Matched clusters must represent at least the 85% of the total computation time.
- Stability = 3 equivalent clusterings "in-a-row".
- Requisites are gradually lowered if can not be matched



17

### SPECFEM3D (64 tasks)



### **Clusters distribution**





### **CPI STACK model**







### SPECFEM3D (64 tasks)

#### Other statistics

Architer   Category Metric Description Cluster 1 Cluster 2   Performance % Duration 0.70236 0.20204   Performance Karchiter Architer   Performance % Duration (µs) 0.70236 0.20204   Performance Total preemted time (µs) 31394.19 13759.21   Performance % preempted time (µs) 31394.19 14559.30   Performance KPC 0.705 0.066   Performance IPC 0.705 0.066   Performance MIPs 1.702.70 1501.76   Performance Memory instructions per second 1.024.80 0.0168   Performance HW floating point instructions per cycle 0.203.02 0.1618   Performance HW floating point instructions rate 656.96 0.381.16   Performance Local L 2 load bandwidth per proc					Archited
CategoryMetric DescriptionCluster 1Cluster 2Performance% Duration (µs)0.70230.26024PerformanceArg. Burst Duration (µs)2.142.664.27947.044.53PerformanceTotal preemted time (µs)31394.1913759.21Performance% preempted time (µs)31394.191.453%PerformancePC0.750.66PerformanceIPC0.750.66PerformanceCPI1.702.771501.75PerformanceMiPs1702.771501.75PerformanceMemory instructions per second1324.801740.67PerformanceHW floating point instructions per cycle0.0200.168PerformanceHW floating point instructions rate656.95381.15PerformanceComputation intensity2.0041.128PerformanceLocal L2 load bandwidth per processor (MB/s)5.634.582.409.52Performance% Loads from local L2 per cycle2.033*0.872*					Archited
CategoryMetric DescriptionCluster 1Cluster 2Performance% Duration0.702360.26024PerformanceAvg. Burst Duration (µs)2.142.664.27947.044.53PerformanceTotal preemted time (µs)31394.1913759.21Performance% preempted time (µs)31394.1913759.21PerformanceIPC0.0750.666PerformanceIPC0.7750.666PerformanceCPI1.012.131.151PerformanceMIPs1.702.701501.75PerformanceMemory instructions per second1.324.801.740.67PerformanceHW floating point instructions per cycle0.02900.0168PerformanceFlop rate (MFLOPs)1.421.36555.95PerformanceComputation intensity2.0041.128PerformanceComputation intensity2.0041.128PerformanceKoads from local L2 per cycle2.039%0.872%					Archite
CategoryMetric DescriptionCluster 1Cluster 2ArchiterPerformance% Duration0.702360.26024PerformanceAvg. Burst Duration (µs)2.142.664,27947.044,53PerformanceTotal preemted time (µs)31394,1913759,21Performance% preempted time (µs)31394,1913759,21Performance% preempted time1,465%1,453%PerformanceIPC0,750,66PerformanceCPI1,0331,51PerformanceMIPs1702,701501,75PerformanceMem.BW (MB/s)260,22203,84PerformanceMemory instructions per second1324,801740,67PerformanceHW floating point instructions per cycle0,2900,168PerformanceFlop rate (MFLOPs)1.421,36555,95PerformanceComputation intensity2,0041,128PerformanceLocal L2 load bandwidth per processor (MB/s)5.634,582.409,52Performance% Loads from local L2 per cycle2,03%0,872%					Archite
Performance   % Duration   0.70236   0.26024     Performance   Avg. Burst Duration (µs)   2.142.664,27   947.044,53     Performance   Total preemted time (µs)   31394.19   13759,21     Performance   % preempted time (µs)   1,465%   1,453%     Performance   IPC   0,75   0,66     Performance   CPI   1,33   1,51     Performance   MIPs   1702,70   1501,75     Performance   Mem.BW (MB/s)   260,22   203,84     Performance   Memory instructions per second   1324,80   1740,67     Performance   HW floating point instructions per cycle   0,209   0,168     Performance   HW floating point instructions rate   656,65   381,15     Performance   MW floating point instructions rate   656,65   381,15     Performance   Computation intensity   2,004   1,128     Performance   Local L2 load bandwidth per processor (MB/s)   5.634,58   2.409,52     Performance   % Loads from local L2 per cycle   2,039%   0,872%	Category	Metric Description	Cluster 1	Cluster 2	Archited
PerformanceAvg. Burst Duration (µs)2.142.664,27947.044,53PerformanceTotal preemted time (µs)31394,1913759,21Performance% preempted time1,465%1,453%PerformanceIPC0,750,66PerformanceCPI1,331,51PerformanceMIPs1702,701501,75PerformanceMem.BW (MB/s)260,22203,84PerformanceMemory instructions per second1324,801740,67PerformanceHW floating point instructions per cycle0,2900,168PerformanceHW floating point instructions rate656,95381,15PerformanceComputation intensity2,0041,128PerformanceLocal L2 load bandwidth per processor (MB/s)5.634,582.409,52Performance% Loads from local L2 per cycle2,03%0,872%	Performance	% Duration	0.70236	0.26024	
PerformanceTotal preemted time (µs)31394,1913759,21Performance% preempted time1,465%1,453%PerformanceIPC0,750,66PerformanceCPI1,331,51PerformanceMIPs1702,701501,75PerformanceMem.BW (MB/s)260,22203,84PerformanceMemory instructions per second1324,801740,67PerformanceHW floating point instructions per cycle0,2000,168PerformanceFlop rate (MFLOPs)1.421,36555,95PerformanceHW floating point instructions rate656,95381,15PerformanceComputation intensity2,0041,128PerformanceLocal L2 load bandwidth per processor (MB/s)5.634,582.409,52Performance% Loads from local L2 per cycle2,039%0,872%	Performance	Avg. Burst Duration (μs)	2.142.664,27	947.044,53	
Performance% preempted time1,465%1,453%PerformanceIPC0,750,66PerformanceCPI1,331,51PerformanceMIPs1702,701501,75PerformanceMem.BW (MB/s)260,22203,84PerformanceMemory instructions per second1324,801740,67PerformanceHW floating point instructions per cycle0,2000,168PerformanceFlop rate (MFLOPs)1.421,36555,95PerformanceHW floating point instructions rate656,95381,15PerformanceComputation intensity2,0041,128PerformanceLocal L2 load bandwidth per processor (MB/s)5.634,582.409,52Performance% Loads from local L2 per cycle2,039%0,872%	Performance	Total preemted time (µs)	31394,19	13759,21	
PerformanceIPC0,66PerformanceCPI1,331,51PerformanceMIPs1702,701501,75PerformanceMem.BW (MB/s)260,22203,84PerformanceMemory instructions per second1324,801740,67PerformanceHW floating point instructions per cycle0,2090,168PerformanceFlop rate (MFLOPs)1.421,36555,95PerformanceHW floating point instructions rate656,95381,15PerformanceComputation intensity2,0041,128PerformanceLocal L2 load bandwidth per processor (MB/s)5.634,582.409,52Performance% Loads from local L2 per cycle2,039%0,872%	Performance	% preempted time	1,465%	1,453%	
PerformanceCPI1,331,51PerformanceMIPs1702,701501,75PerformanceMem.BW (MB/s)260,22203,84PerformanceMemory instructions per second1324,801740,67PerformanceHW floating point instructions per cycle0,2000,168PerformanceFlop rate (MFLOPs)1.421,36555,95PerformanceHW floating point instructions rate656,95381,15PerformanceComputation intensity2,0041,128PerformanceLocal L2 load bandwidth per processor (MB/s)5.634,582.409,52Performance% Loads from local L2 per cycle2,039%0,872%	Performance	IPC	0,75	0,66	
PerformanceMIPs1702,701501,75PerformanceMem.BW (MB/s)260,22203,84PerformanceMemory instructions per second1324,801740,67PerformanceHW floating point instructions per cycle0,2900,168PerformanceFlop rate (MFLOPs)1.421,36555,95PerformanceHW floating point instructions rate656,95381,15PerformanceComputation intensity2,0041,128PerformanceLocal L2 load bandwidth per processor (MB/s)5.634,582.409,52Performance% Loads from local L2 per cycle2,039%0,872%	Performance	СРІ	1,33	1,51	
PerformanceMem.BW (MB/s)260,22203,84PerformanceMemory instructions per second1324,801740,67PerformanceHW floating point instructions per cycle0,2900,168PerformanceFlop rate (MFLOPs)1.421,36555,95PerformanceHW floating point instructions rate656,95381,15PerformanceComputation intensity2,0041,128PerformanceLocal L2 load bandwidth per processor (MB/s)5.634,582.409,52Performance% Loads from local L2 per cycle2,039%0,872%	Performance	MIPs	1702,70	1501,75	
PerformanceMemory instructions per second1324,801740,67PerformanceHW floating point instructions per cycle0,2900,168PerformanceFlop rate (MFLOPs)1.421,36555,95PerformanceHW floating point instructions rate656,95381,15PerformanceComputation intensity2,0041,128PerformanceLocal L2 load bandwidth per processor (MB/s)5.634,582.409,52Performance% Loads from local L2 per cycle2,039%0,872%	Performance	Mem.BW (MB/s)	260,22	203,84	
PerformanceHW floating point instructions per cycle0,2900,168PerformanceFlop rate (MFLOPs)1.421,36555,95PerformanceHW floating point instructions rate656,95381,15PerformanceComputation intensity2,0041,128PerformanceLocal L2 load bandwidth per processor (MB/s)5.634,582.409,52Performance% Loads from local L2 per cycle2,0390,872%	Performance	Memory instructions per second	1324,80	1740,67	
PerformanceFlop rate (MFLOPs)1.421,36555,95PerformanceHW floating point instructions rate656,95381,15PerformanceComputation intensity2,0041,128PerformanceLocal L2 load bandwidth per processor (MB/s)5.634,582.409,52Performance% Loads from local L2 per cycle2,039%0,872%	Performance	HW floating point instructions per cycle	0,290	0,168	
PerformanceHW floating point instructions rate656,95381,15PerformanceComputation intensity2,0041,128PerformanceLocal L2 load bandwidth per processor (MB/s)5.634,582.409,52Performance% Loads from local L2 per cycle2,039%0,872%	Performance	Flop rate (MFLOPs)	1.421,36	555,95	
PerformanceComputation intensity2,0041,128PerformanceLocal L2 load bandwidth per processor (MB/s)5.634,582.409,52Performance% Loads from local L2 per cycle2,039%0,872%	Performance	HW floating point instructions rate	656,95	381,15	
PerformanceLocal L2 load bandwidth per processor (MB/s)5.634,582.409,52Performance% Loads from local L2 per cycle2,039%0,872%	Performance	Computation intensity	2,004	1,128	
Performance% Loads from local L2 per cycle2,039%0,872%	Performance	Local L2 load bandwidth per processor (MB/s)	5.634,58	2.409,52	
	Performance	% Loads from local L2 per cycle	2,039%	0,872%	

Category	Metric Description	Cluster 1	Cluster 2
Architecture	% Instr. Completed	32,25%	32,22%
Architecture	L1 misses per Kinstr.	41,37	34,82
Architecture	L2 misses per Kinstr.	1,194	1,060
Architecture	Bytes from maim memory per floating point instruction finished	1,649	1,361
Architecture	Number of Loads per Load miss	24,81	66,68
Architecture	Number of Stores per Store miss	6,74	5,27
Architecture	Number of Loads&Stores per L1 miss	18,81	33,28
Architecture	L1 cache hit rate	94,68%	97,00%
Architecture	Number of Loads per (D)TLB miss	11.403,59	5.873,04
Architecture	Number of Loads&Stores per (D)TLB miss	12.945,98	6.425,90
Architecture	% TLB misses per cycle	0,005%	0,012%
Architecture	Total Loads from local L2 (M) (total_Id_I_L2)	94,320	17,828
Architecture	Local L2 load traffic (MB)	12.073,007	2.281,926

Category	Metric Description	Cluster 1	Cluster 2
Instruction Mix	FMA ops per floating point instruction	0,763	0,633
Instruction Mix	Instructions per Load/Store	1,285	0,863
Instruction Mix	HW floating point instructions (flips)	1.407.628.943	360.968.332
Instruction Mix	Total floating point operations (flops)	3.045.492.593	526.507.674
Instruction Mix	Total FP Load&Store operations (fp_tot_ls)	1.519.651.784	466.816.330
Instruction Mix	FMA %	81,04%	82,15%
Instruction Mix	Memory Mix	25,10%	37,35%
Instruction Mix	Load Mix	22,11%	34,14%
Instruction Mix	Store Mix	2,99%	3,21%
Instruction Mix	FPU Mix	12,44%	8,18%
Instruction Mix	FXU Mix	3,93%	12,64%





### SPECMPI'07 ZEUSMP2 (128 tasks)



196596351.58 us 197795090.48 us

128 tasks, 200 Mb









195397612,67 us

1



200192568,30 us

198993829,39 us

### SPECMPI'07 – LESLIE3D (256 tasks)







### **GROMACS (64 tasks)**











### Conclusion



- We can get out much mode information than we currently do
- Can be done on-line
- Detailed analysis and visualization to find out how



