Paraver Advanced

Jesús Labarta, Judit Gimenez
Jordi Caubet, Francesc Escale

CEPBA-UPC

Technology Transfer | Research | Training | Mobility of Researchers | User Support | Education | HPC Facilities | Parallel Expertise

Index

- **Case studies**
  - OS interrupts
  - On the effect of MPI environment variables
    - ✓ CSS
    - ✓ MP_SHARED_MEMORY
    - ✓ MP_EAGER_LIMIT
    - ✓ US vs IP
  - Sweep3D
    - ✓ MPI + OpenMP
    - ✓ Computation phase analysis
  - SPEC OMP2001
    - ✓ Load Balance & invalidations @ ammp
    - ✓ Memory bandwidth @ swim

OS internals

Jesús Labarta, Judit Gimenez, Jordi Caubet

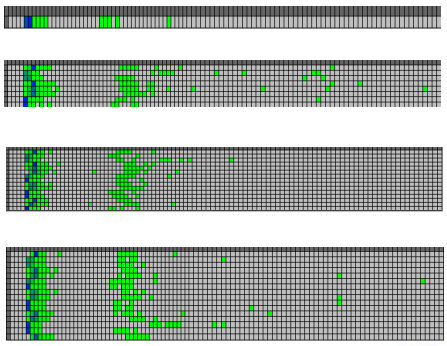
CEPBA-UPC

Technology Transfer | Research | Training | Mobility of Researchers | User Support | Education | HPC Facilities | Parallel Expertise

Tracing overhead

```
DO I=1,10000
 call ompitrace_eventandcounters(1000,I)
 call ompitrace_eventandcounters(1000,0)
END DO
```

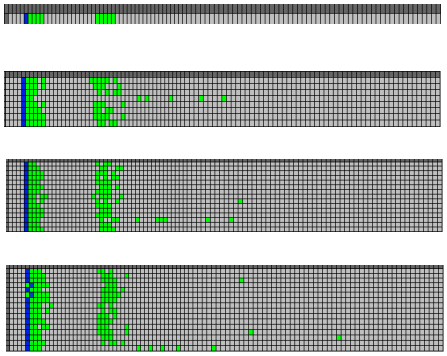
- **OpenMP**
  - 0 – 100 µs



Tracing overhead

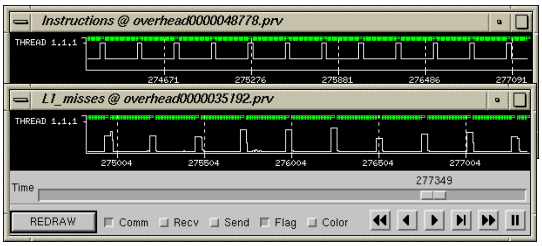
```
DO I=1,10000
 call ompitrace_eventandcounters(1000,I)
 call ompitrace_eventandcounters(1000,0)
END DO
```

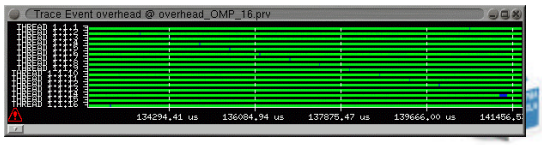
- **MPI**
  - 0 – 100 µs



System interrupts

- **Observable**
  - Per process kernel lock in access to hardware counters information
  - Quantum
  - Cost
  - Skew between CPUs





On the effect of environment variables

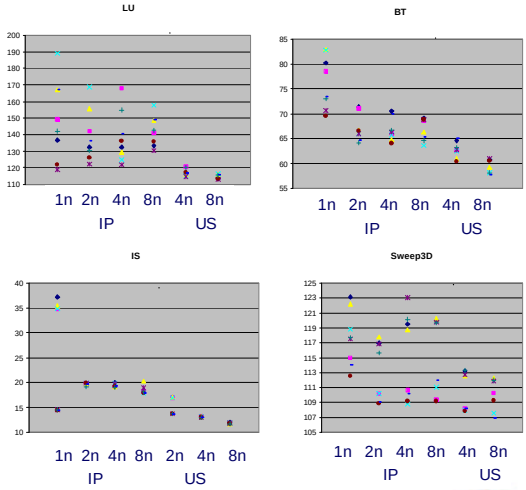
Jesús Labarta, Judit Gimenez, Jordi Caubet

CEPBA-UPC

Technology Transfer | Research | Training | Mobility of Researchers | User Support | Education | HPC Facilities | Parallel Expertise

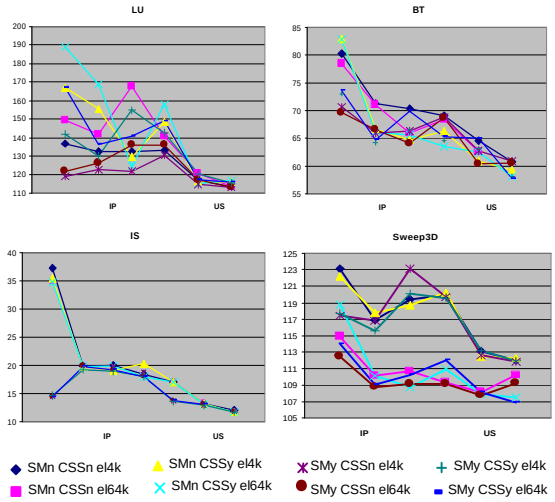
Effect of environment variables

- **16 tasks**
- **Environment**
  - MP_SHARED_MEMORY
  - MP_EAGER_LIMIT
    - ✓ 4k / 64k
  - MP_CSS_INTERRUPT
  - IP / US
  - Number of nodes
    - ✓ 1, 2, 4, 8



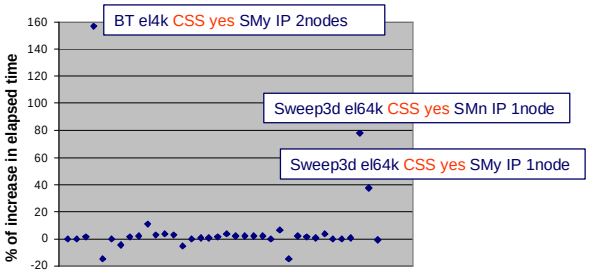
Effect of environment variables

- **Often <10%, sometimes significant**
- **Good combination:**
  - unpredictable
  - Autonomic Computing needed here too
- **US reduces influence of others**



Tracing overhead

- **Comparison of elapsed time when tracing vs. un-instrumented**
  - Generally small difference (<5%)
  - CSS responsible of outliers

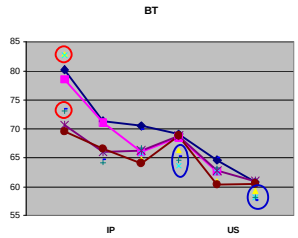


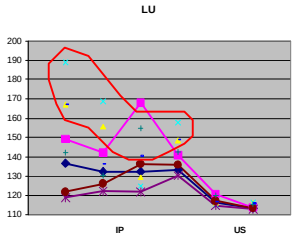
Studies

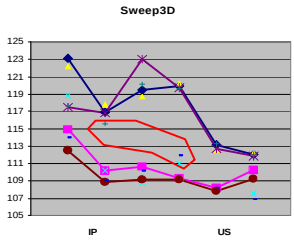
- Effect of MP_SHARED_MEMORY
- Effect of MP_EAGER_LIMIT
- Comparison between US and IP
- MP_CSS_INTERRUPT
- 2 nodes better than 1 node
- US vs. SM
- …
- …
- ……8

CSS

- Usually bad
- May be good



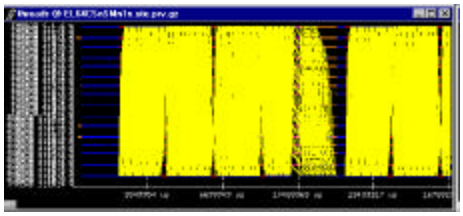




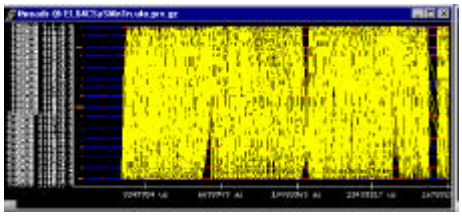
CSS: ... sweep3d

- **The bad**
- **OS scheduling**
  - Conflict between application and MPI implementation threads
- **Ute traces ? Paraver**

CSS no



Sweep3d, SMy, 1 node

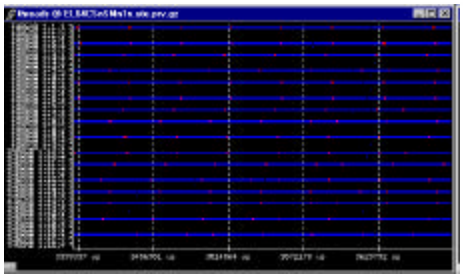


CSS yes

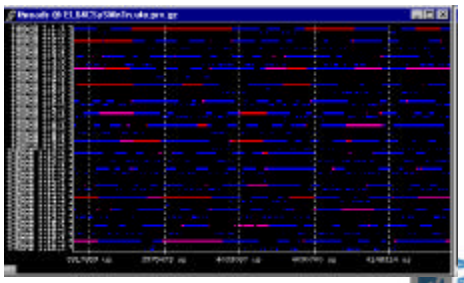
CSS: ... sweep3d

- **The bad**
- **OS scheduling**
  - Is yielding used?

CSS no



CSS yes

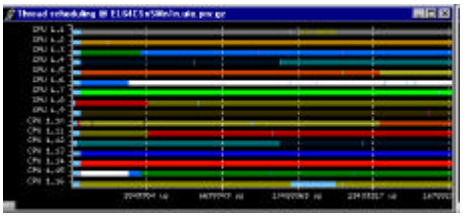


CSS: ... sweep3d

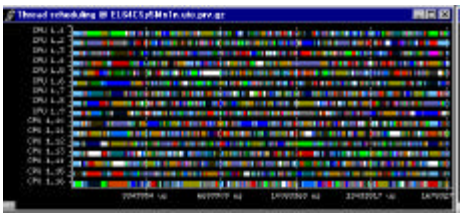
- **The bad**
  - Process migrations
  - Can internal threads be bound to their corresponding master thread?

Time span : ~ 6 s

CSS no



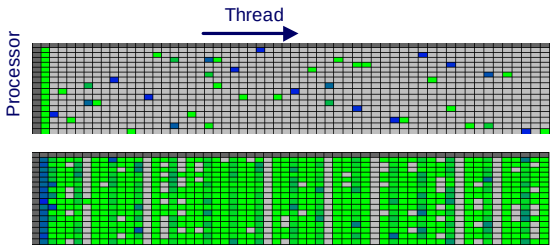
CSS yes



Time span : ~ 12 s

CSS: ... sweep3d

- **Time spent by each thread on each processor**
  - Visible: Thread to processor mapping if CSS ==no
  - Curious: Single thread taking more % of CPUtime at each processor if CSS == yes: Idle processor (15%)

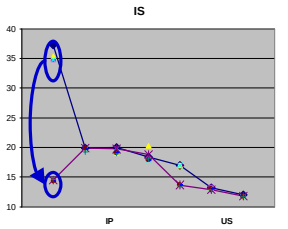


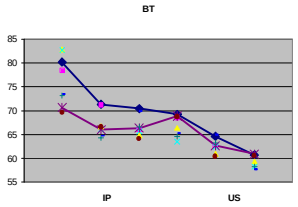
CSS no

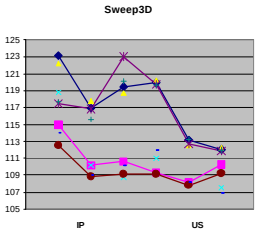
CSS yes

MP_SHARED_MEMORY

- **Important difference**
- **Or not**
- **Generally worthwhile**
  - Default NO ??





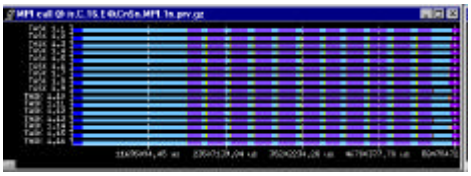


◆ SMn CSSn el4k ✳ SMy CSSn el4k
■ SMn CSSn el64k ● SMy CSSn el64k

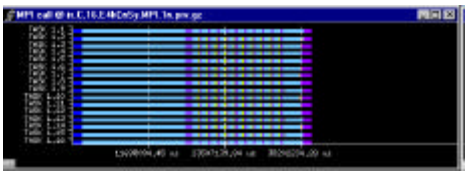
MP_SHARED_MEMORY: ... IS

- **Load MPI_call configuration**
- **Important difference**
  - Need to quantify
- **Shared memory: first time slower**
  - Significant influence in average

SM no



IS, CSSn, el4k

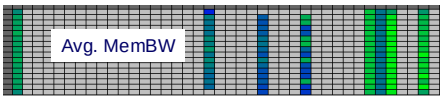


SM yes

MP_SHARED_MEMORY: ... IS

- **2D analysis configuration file**
- **A given environment set up may be good for one call and not so good for other**
- **Statistics**
  - Care with selected area
  - May be misleading
    - ✓ Low miss ratio not always synonym of good

MPI call

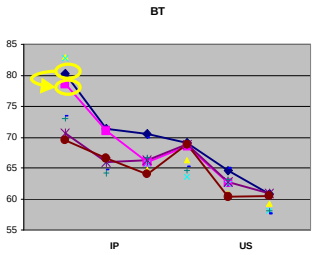


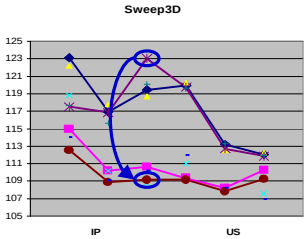
	SM yes	SM no	Ratio
time (s)	14.46	37.25	0.39
MPI_Alltoallv duration (us)	441314	2488005	0.18
MPI_Alltoall duration (us)	520	8334	0.06
MPI_Allreduce duration (us)	205017	186966	1.10

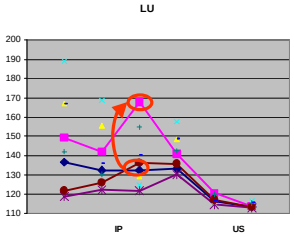
Avg. Values @ MPI_Alltoallv	SM yes	SM no	Ratio
Stores	9421214	29634927	0.32
Loads	11632523	66742482	0.17
Mem Ops mix (%)	74	43.6	1.70
L2 miss ratio (%)	4.1	1.6	2.56
Mem BW (MB/s)	375.8	94.2	3.99
CPU BW (MB/s)	440	310	1.42
IPC	0.2	0.24	0.83

MP_EAGER_LIMIT

- **Increasing eager limit we expect**
  - Not much difference
  - An improvement
- **but may get bad !!!**



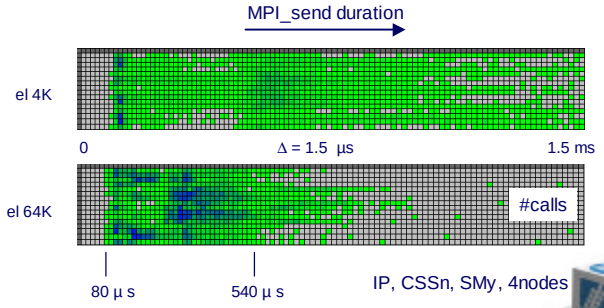




◆ SMn CSSn el4k ✳ SMy CSSn el4k
■ SMn CSSn el64k ● SMy CSSn el64k

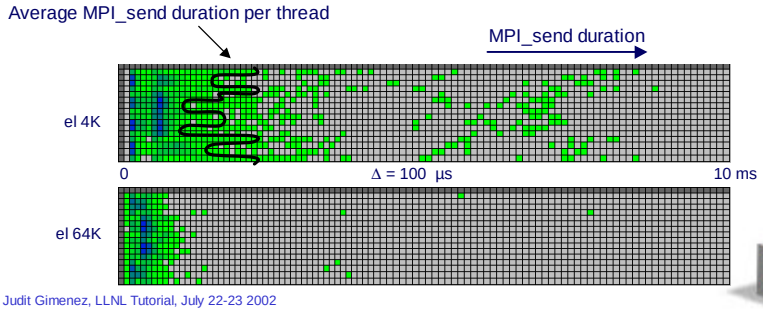
MP_EAGER_LIMIT: … sweep3d

- **Expectation:**
  - Sufficient Eager limit (64K) ➔ sends do not wait
- **Time distribution of MPI_send**
  - Fast calls approximately equal
  - Slow calls faster



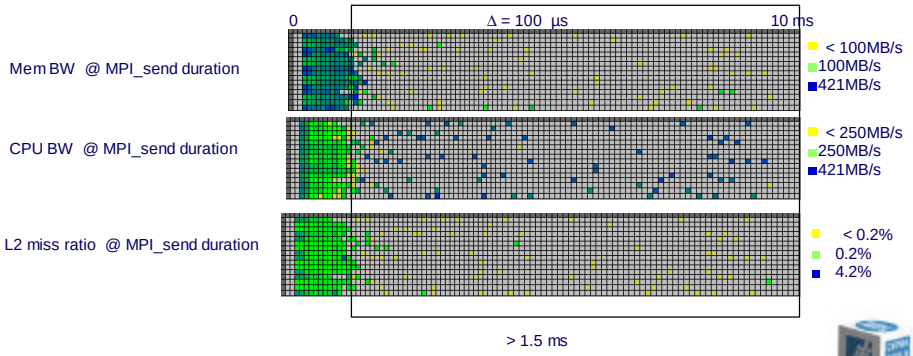
MP_EAGER_LIMIT: … sweep3d

- **A view from the heights**
  - Some very slow sends
  - Not uniform distribution
    - ✓ Time distribution
    - ✓ Per thread distribution



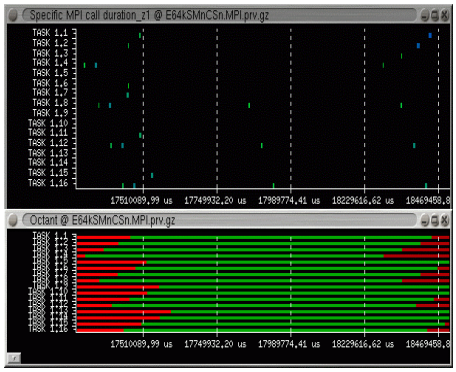
MP_EAGER_LIMIT: … sweep3d

- **Some very long sends even with large eager limit**
  - Configuration 1node …..



MP_EAGER_LIMIT: … sweep3d

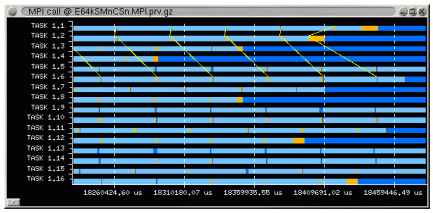
- **Identification**
  - Few cases
    - ✓ difficult based on direct navigation state timeline
  - Specific view
    - ✓ MPI sends of duration >1.5ms

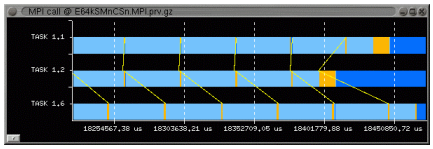


- **Correlation**
  - With source
  - With iteration

MP_EAGER_LIMIT: ... sweep3d

- **Focus View**
  - Filter Displayed communications
  - Hide non relevant threads
- **Interpretation**
  - MP_EAGER_LIMIT ??
    - ✓ Several consecutive sends
    - ✓ What if previous one not yet received ?



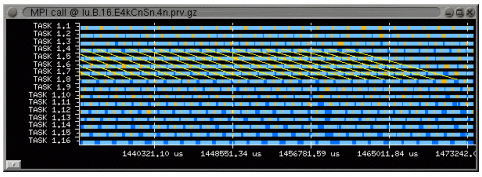


MP_EAGER_LIMIT: ... LU

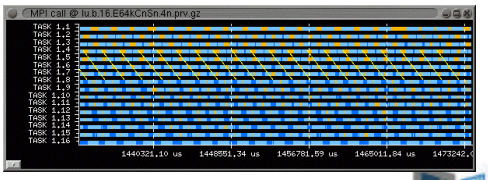
- **Increasing eager limit is bad !!!**
  - For some calls
  - Even if all messages <4K
  - Flow control issue

	4K	64K	Ratio
time (s)	136.20	167.90	0.81
MPI_send duration (us)	162.4	255.8	0.63
MPI_recv duration (us)	225.6	273.2	0.83
MPI_Irecv duration (us)	54.4	54.6	1.00
MPI_Wait duration (us)	7260	4544	1.60

4K



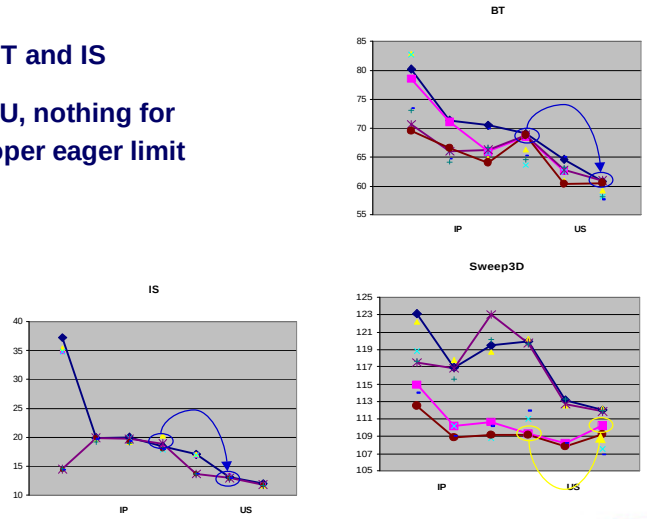
64K



CSS no, SM no, 4 nodes

US vs IP

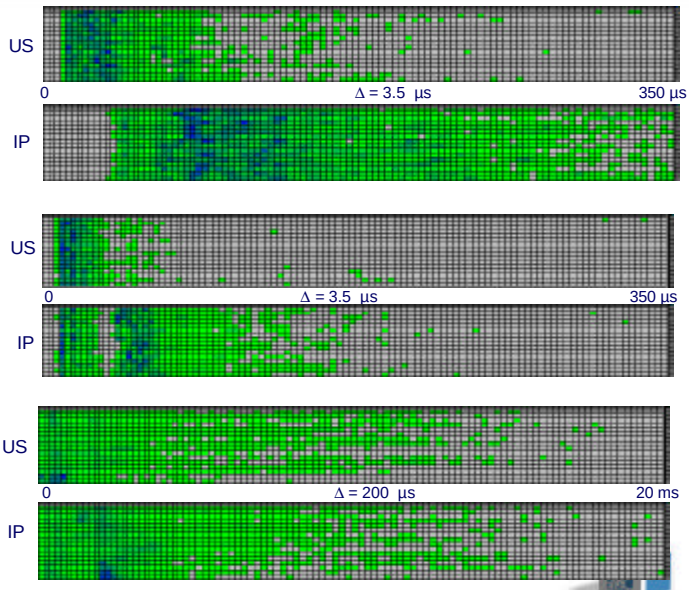
- **Improves for BT and IS**
- **Very little for LU, nothing for Sweep3d if proper eager limit**



◆ SMn CSSn el4k ✳ SMy CSSn el4k
■ SMn CSSn el64k ● SMy CSSn el64k

US vs IP: … BT

- **Isend**
  - Speed
  - US less variance
- **Irecv**
  - Uni/Bi modal
- **Wait**
  - Variance



BT, CSSn, SMn, e4k, 8 nodes

Sweep3D analysis

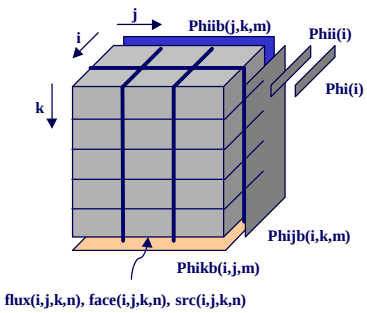
Jesús Labarta, Judit Gimenez

CEPBA-UPC

Technology Transfer | Research | Training | Mobility of Researchers | User Support | Education | HPC Facilities | Parallel Expertise

MPI Parallelization

- **Grid Partitioning of `ij` plane**
- **One column per MPI task**
- **Reduction on each direction:**
  - Communicate on `i` and `j`
  - Pipeline `k` dimension
  - + overlap consecutive sweeps if possible



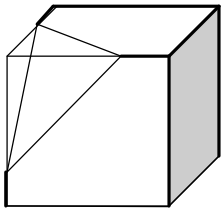
- **Input file parameters**
  - Blocks on `i` and `j` directions
  - Pipelining block

MPI data parallelization

Flux, face,src:	DISTRIBUTED
phii, phi:	PRIVATIZED
phikb:	DISTRIBUTED
phijb,phiib:	DISTRIBUTED&REPLICATED => Communication

OpenMP Parallelization

- **Thought for mixed mode (MPI + OpenMP)**
  - Just core computational loop
- **Reduction on each direction**
  - Compute along diagonal wavefronts



```
DO idiag
   DO jkm =1, #points in wavefront
      j,k,m=f(idiag,jkm)
      DO n,i      ! phi, src
      DO i        ! phijb, phikb, phii, phi
      DO n,i      ! flux, phi
      DO i        ! face, phii, phijb, phikb
```

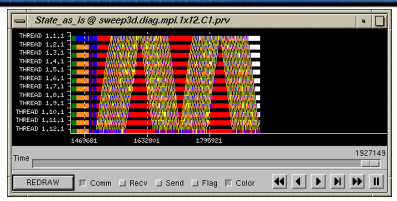

MPI vs. OpenMP: some numbers

- **Problem**
  - size: 50^3
  - k plane pipelining: 10

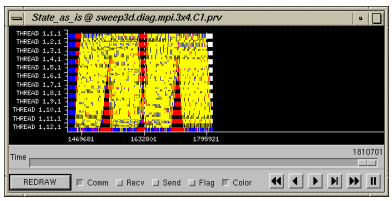
NB Domains	OpenMP time	Decomposition	MPI time
6	7.78	1x6	3.97
		2x3	3.61
		3x2	3.71
		6x1	4.47
12	6.55	1x12	3.50
		2x6	2.74
		3x4	2.21
		4x3	2.25
		6x2	2.97
		12x1	3.98

MPI : decomposition effect

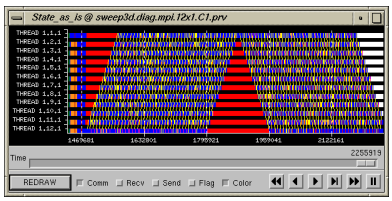
- 1 x 12



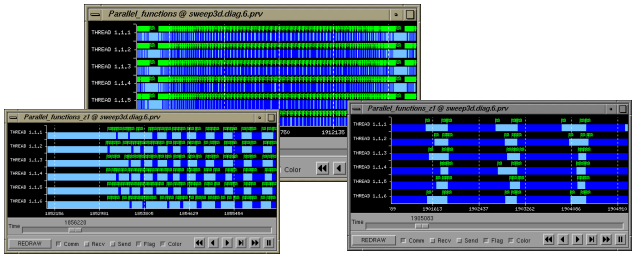
- 3 x 4



- 12 x 1



Pure OpenMP




```
DO idiag
    :
   DO jkm =1, #points in wavefront
      j,k,m=f(idiag,jkm)
      DO n,i      ! phi, src
      DO i        ! phijb, phikb, phii, phi
      DO n,i      ! flux, phi
      DO i        ! face, phii, phijb, phikb
```

Parallel

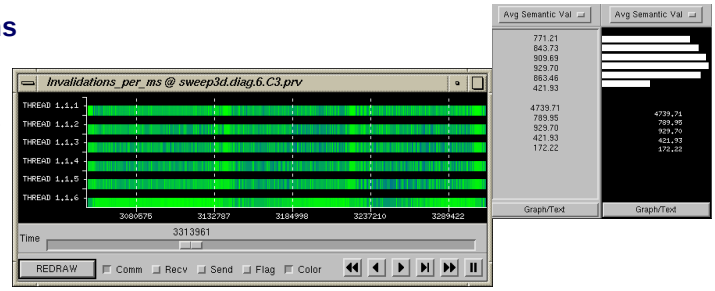
Computation:
- **Complex**
- **Overhead**

Triangular trip count
- **OpenMP RTL overhead**
- **Invalidation traffic**

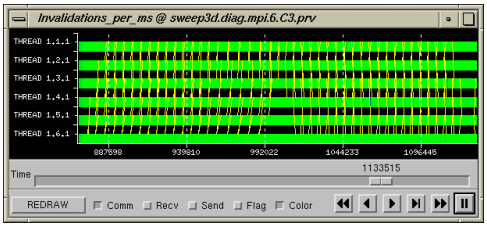
Computation:
- **Complex**
- **Overhead**

Pure OpenMP

- **Invalidations**
  - OpenMP



  - MPI



Alternative structures

```
DO idiag
   DO jkm =1, #points in wavefront
      j,k,m=f(idiag,jkm)
      DO n,i      ! phi, src
      DO i        ! phijb, phikb, phii, phi
      DO n,i      ! flux, phi
      DO i        ! face, phii, phijb, phikb
```

mkji form

in the distributed source

```
DO m
   DO k
      DO j
         DO n,i      ! phi, src
         DO i        ! phijb, phikb, phii, phi
         DO n,i      ! flux, phi
         DO i        ! face, phii, phijb, phikb
```

Alternative forms

- **mkji**

```
DO k
   DO j
      DO m
         DO n,i
         DO i
         DO n,i
         DO i
```

- **Kjkmi**

```
DO K
   DO j
      DO k
         DO m
            DO n,i
            DO i
            DO n,i
            DO i
```

- **kjmi**

```
DO k
   DO j
      DO m
         DO n,i
         DO i
         DO n,i
         DO i
```

- **jkmi**

```
DO j
   DO k
      DO m
         DO n,i
         DO i
         DO n,i
         DO i
```

- . . .

OpenMP: some numbers

- **Problem**
  - size: 50^3
  - k plane pipelining: 10

Insufficient parallelism

Locality

Contention on locks

Invalidation, RTL overhead

Elapsed Time	OMP_NUM_THREADS																
Version	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
ccrit	24.40	28.26	24.41	26.84	26.47	29.28	30.34						30.43				
cpipe	24.60	25.63	18.45	13.01	12.53	10.06	7.67						7.76				
diag	16.21	17.28	13.09	11.40	9.64	8.50	7.78						6.55				
kjmi	12.95	14.86	10.01	7.35	5.82	4.89	4.34	3.80	3.62	3.38	3.09	3.04	2.88	2.69	2.69	2.64	2.53
Kjkmi	12.94	14.91	8.47	6.35	4.91	4.24	3.58	3.46	2.90	2.81	2.78	2.65	2.29	2.22	2.16	2.19	2.15

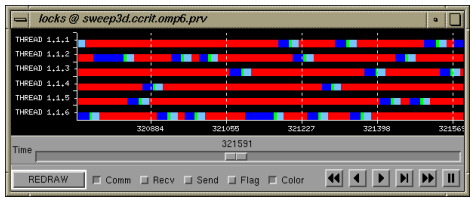
Invalidations

Instruction count overhead?

OpenMP : contention on locks

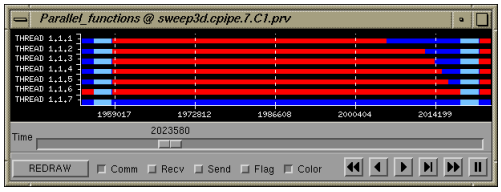
- **Version ccrit, 6 Threads**

```
DO k                                  Parallel
   DO j
      DO m
         DO n,i
         DO i
         DO n,i                       CRITICAL
         DO i
```

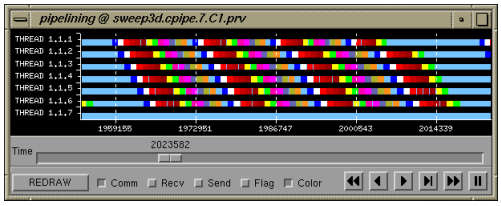


OpenMP: Insufficient parallelism

- **Version cpipe, 7 threads**
  - outer iteration count: 6
  - parallel + worksharing



  - Internal pipelined iteration



Comparing modes

Single mode better

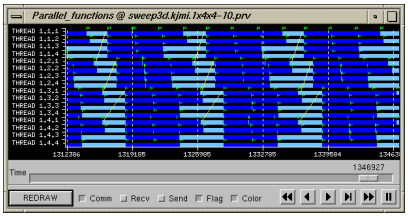
		Seq	OpenMP				
MPI	partition		1	2	4	8	16
0	-				66,33	37,39	22,77
1	1x1				66,99	38,21	23,55
2	1x2			79,34	50,90	36,58	
	2x1			89,60	57,52	42,36	
4	1x4	66,20	71,67	41,92	28,24		
	2x2	79,72	86,71	52,01	34,60		
	4x1	76,04	86,04	54,51	37,05		
8	1x8	37,83	40,27	24,90			
	2x4	42,93	46,04	28,84			
	4x2	46,39	50,74	33,79			
	8x1	44,71	50,91	36,34			
16	1x16	23,12	24,82				
	4x4	24,75	26,86				
	16x1	28,40	32,60				

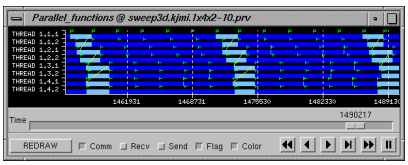
Decomposition effect

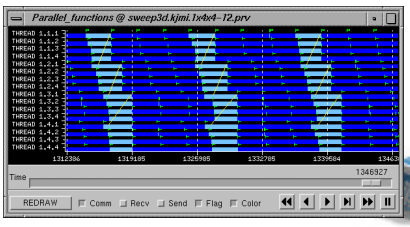
Scheduling Interference

Mixed mode: scheduling interference

- **MPI (4 tasks) + kjmi**
  - 4 threads, k pipeline=10
  - 2 threads, k pipeline=10
  - 4 threads, k pipeline=12
    - ✓ Less K iterations

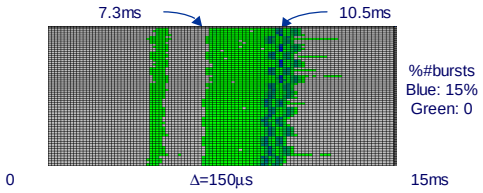






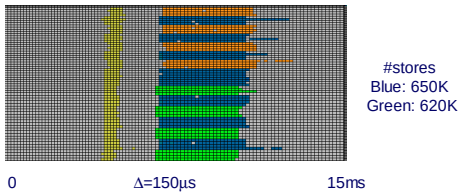
Computation distribution histograms

- **Sweep3d (300 x 300 x 125) , 8x8x1 (MPI_I x MPI_J x OpenMP)**
- **2D**
  - Cw: core loop duration
  - Statistic: % #bursts
- **Comments**
  - Load imbalance ?



Computation distribution histograms

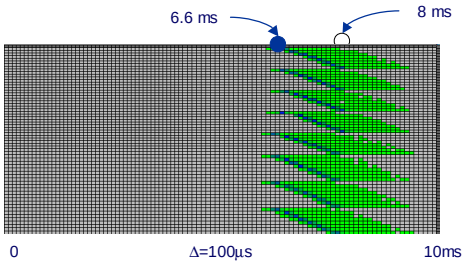
- **8x8x1**
- **2D**
  - Cw: core loop duration
  - Statistic: average value
  - Dw:#stores



- **Comments**
  - Load imbalance: stores = f (CPU)
    - ✓ 300 = 37 * 8 + 4
  - Precision of measurements (consistent detection of small percentual variations)
  - #Stores ≠ f(time) ?

Computation distribution histograms

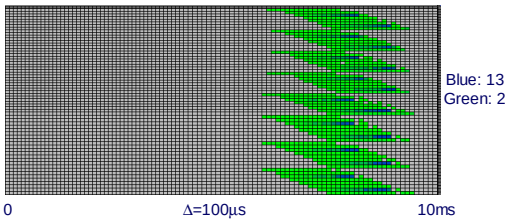
- **Sweep3d (300 x 300 x 125) ,1x8x8 (MPI_I x MPI_J x OpenMP)**
- **2D**
  - Cw: core loop duration
  - Statistic: #bursts



- **Comments**
  - Pipeline
    - ✓ 175 µs

Computation distribution histograms

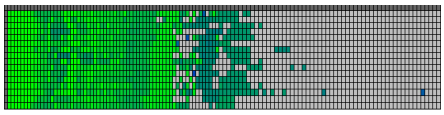
- **1x8x8**
- **2D**
  - Cw: core loop duration
  - Statistic: average value
  - Dw: CPU/memory BW ratio



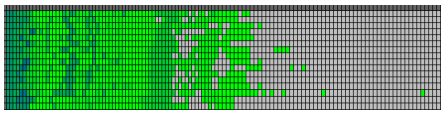
- **Comments**
  - Generally low reuse
  - There is structure
    - ✓ Interpretation ?

Performance indices correlation

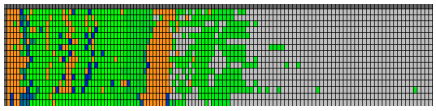
- **Example**
  - L2 misses vs Memory BW



  - IPC vs Memory BW



  Loads vs Memory BW



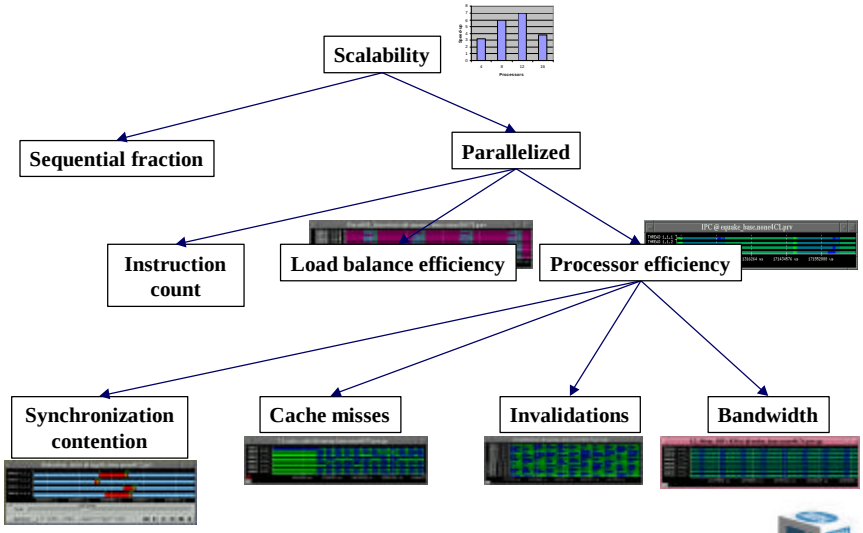
Analysis of the SPEC OMP2001 Benchmarks with Paraver

Jesús Labarta, Judit Gimenez

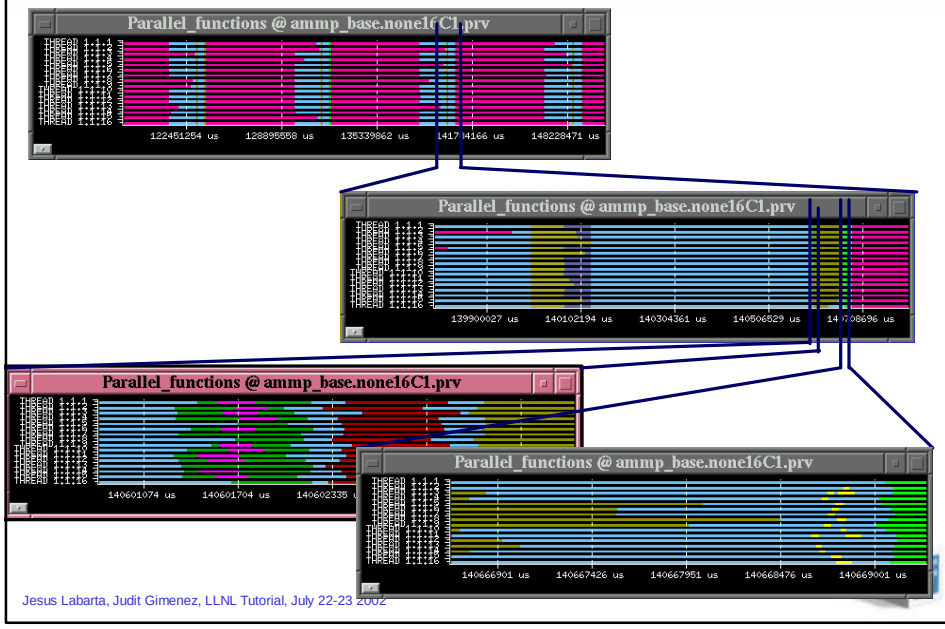
CEPBA-UPC

Technology Transfer | Research | Training | Mobility of Researchers
User Support | Education | HPC Facilities | Parallel Expertise

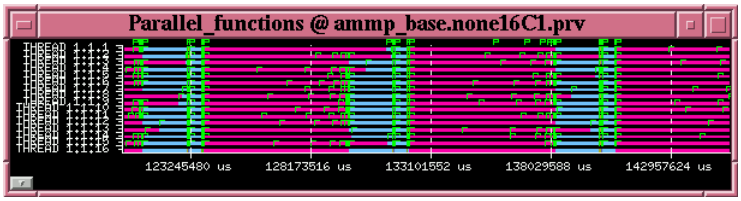
Analysis methodology

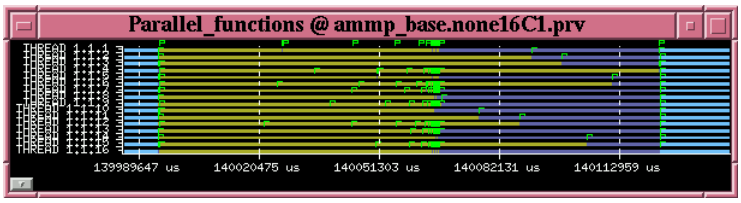


Ammp: load balance

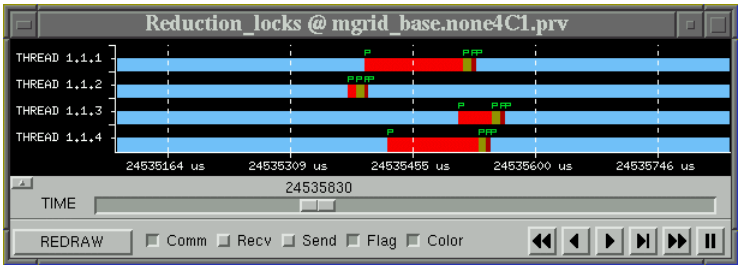


Ammp: loop scheduling

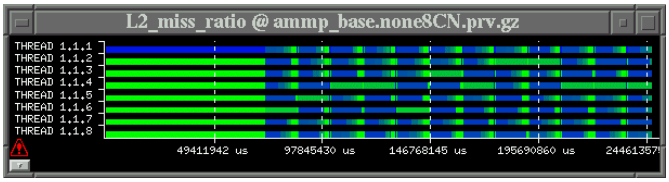


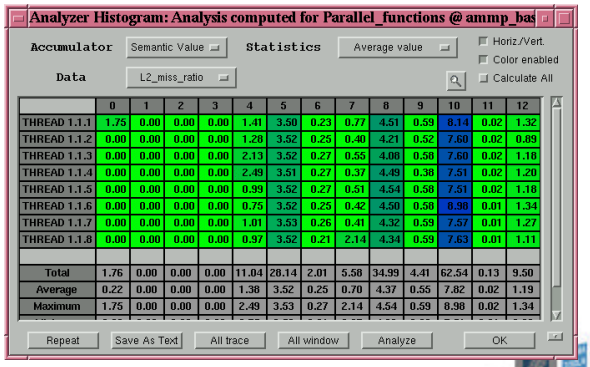


Synchronization



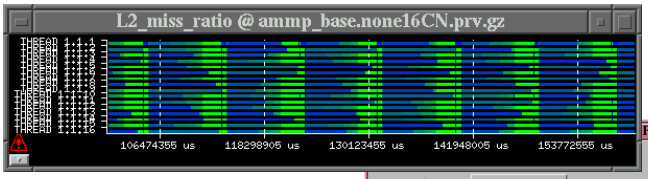
Ammp: L2 miss ratio

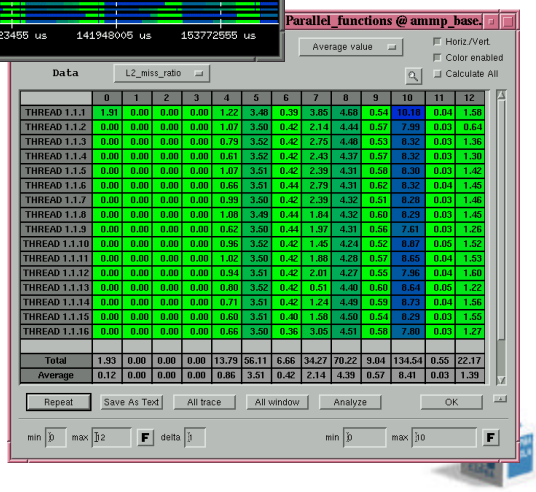




	0	1	2	3	4	5	6	7	8	9	10	11	12
THREAD 1.1.1	1.75	0.00	0.00	0.00	1.41	3.50	0.23	0.77	4.51	0.59	8.14	0.02	1.32
THREAD 1.1.2	0.00	0.00	0.00	0.00	1.28	3.52	0.25	0.40	4.21	0.52	7.60	0.02	0.89
THREAD 1.1.3	0.00	0.00	0.00	0.00	2.13	3.52	0.27	0.55	4.08	0.58	7.60	0.02	1.18
THREAD 1.1.4	0.00	0.00	0.00	0.00	2.49	3.51	0.27	0.37	4.49	0.38	7.51	0.02	1.20
THREAD 1.1.5	0.00	0.00	0.00	0.00	0.99	3.52	0.27	0.51	4.54	0.58	7.51	0.02	1.18
THREAD 1.1.6	0.00	0.00	0.00	0.00	0.75	3.52	0.25	0.42	4.50	0.58	8.98	0.01	1.34
THREAD 1.1.7	0.00	0.00	0.00	0.00	1.01	3.53	0.26	0.41	4.32	0.59	7.57	0.01	1.27
THREAD 1.1.8	0.00	0.00	0.00	0.00	0.97	3.52	0.21	2.14	4.34	0.59	7.63	0.01	1.11
Total	1.76	0.00	0.00	0.00	11.04	28.14	2.01	5.58	34.99	4.41	62.54	0.13	9.50
Average	0.22	0.00	0.00	0.00	1.38	3.52	0.25	0.70	4.37	0.55	7.82	0.02	1.19
Maximum	1.75	0.00	0.00	0.00	2.49	3.53	0.27	2.14	4.54	0.59	8.98	0.02	1.34

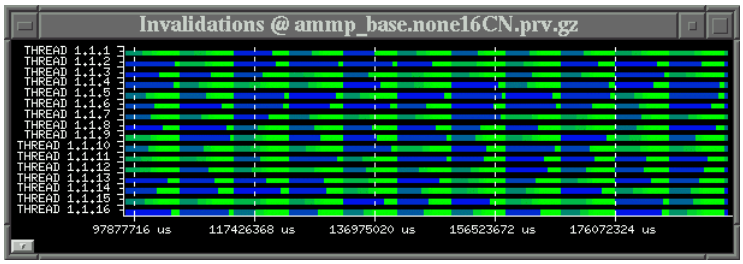
Ammp: L2 miss ratio



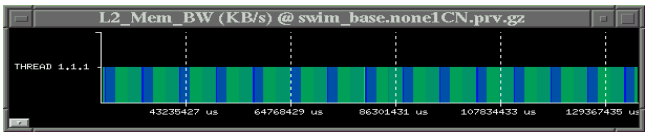


	0	1	2	3	4	5	6	7	8	9	10	11	12
THREAD 1.1.1	1.91	0.00	0.00	0.00	1.22	3.48	0.39	3.85	4.68	0.54	10.18	0.04	1.58
THREAD 1.1.2	0.00	0.00	0.00	0.00	1.07	3.50	0.42	2.14	4.44	0.57	7.99	0.03	0.64
THREAD 1.1.3	0.00	0.00	0.00	0.00	0.79	3.52	0.42	2.75	4.48	0.53	8.32	0.03	1.36
THREAD 1.1.4	0.00	0.00	0.00	0.00	0.61	3.52	0.42	2.43	4.37	0.57	8.32	0.03	1.30
THREAD 1.1.5	0.00	0.00	0.00	0.00	1.07	3.51	0.42	2.39	4.31	0.58	8.30	0.03	1.42
THREAD 1.1.6	0.00	0.00	0.00	0.00	0.66	3.51	0.44	2.79	4.31	0.62	8.32	0.04	1.45
THREAD 1.1.7	0.00	0.00	0.00	0.00	0.99	3.50	0.42	2.39	4.32	0.51	8.28	0.03	1.46
THREAD 1.1.8	0.00	0.00	0.00	0.00	1.08	3.49	0.44	1.84	4.32	0.60	8.29	0.03	1.45
THREAD 1.1.9	0.00	0.00	0.00	0.00	0.62	3.50	0.44	1.97	4.31	0.56	7.61	0.03	1.26
THREAD 1.1.10	0.00	0.00	0.00	0.00	0.96	3.52	0.42	1.45	4.24	0.52	8.87	0.05	1.52
THREAD 1.1.11	0.00	0.00	0.00	0.00	1.02	3.50	0.42	1.88	4.28	0.57	8.65	0.04	1.53
THREAD 1.1.12	0.00	0.00	0.00	0.00	0.94	3.51	0.42	2.01	4.27	0.55	7.96	0.04	1.60
THREAD 1.1.13	0.00	0.00	0.00	0.00	0.80	3.52	0.42	0.51	4.40	0.60	8.64	0.05	1.22
THREAD 1.1.14	0.00	0.00	0.00	0.00	0.71	3.51	0.42	1.24	4.49	0.59	8.73	0.04	1.56
THREAD 1.1.15	0.00	0.00	0.00	0.00	0.60	3.51	0.40	1.50	4.50	0.54	8.29	0.03	1.55
THREAD 1.1.16	0.00	0.00	0.00	0.00	0.66	3.50	0.36	3.05	4.51	0.58	7.80	0.03	1.27
Total	1.93	0.00	0.00	0.00	13.79	56.11	6.66	34.27	70.22	9.04	134.54	0.55	22.17
Average	0.12	0.00	0.00	0.00	0.86	3.51	0.42	2.14	4.39	0.57	8.41	0.03	1.39

Invalidations

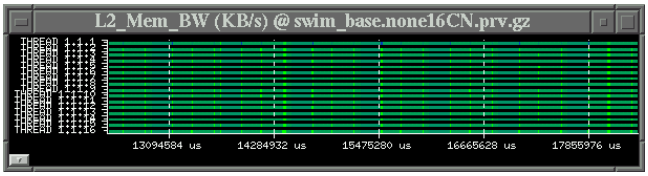


Bandwidth

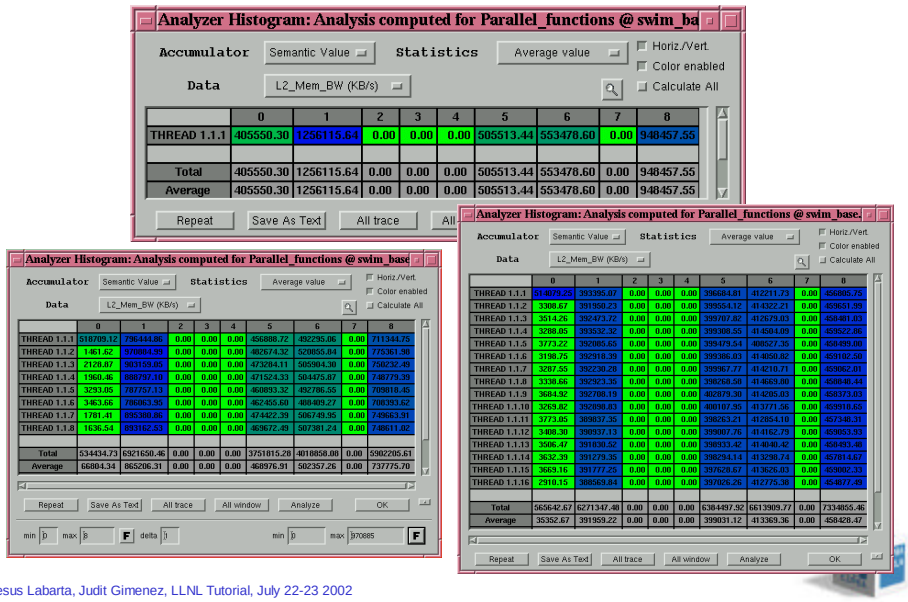


Larger Y scale

Same Y scale



Bandwidth



Jesus Labarta, Judit Gimenez, LLNL Tutorial, July 22-23 2002