Early Results For RAPS8

Bob Carruthers Cray
Deborah Salmond ECMWF
Sami Saarinen ECMWF

Introduction

- **Description of RAPS8**
- **Initial Port**
- **Results for the Forecast Model**
- **Preliminary Results for 4DVAR**

Description of RAPS8

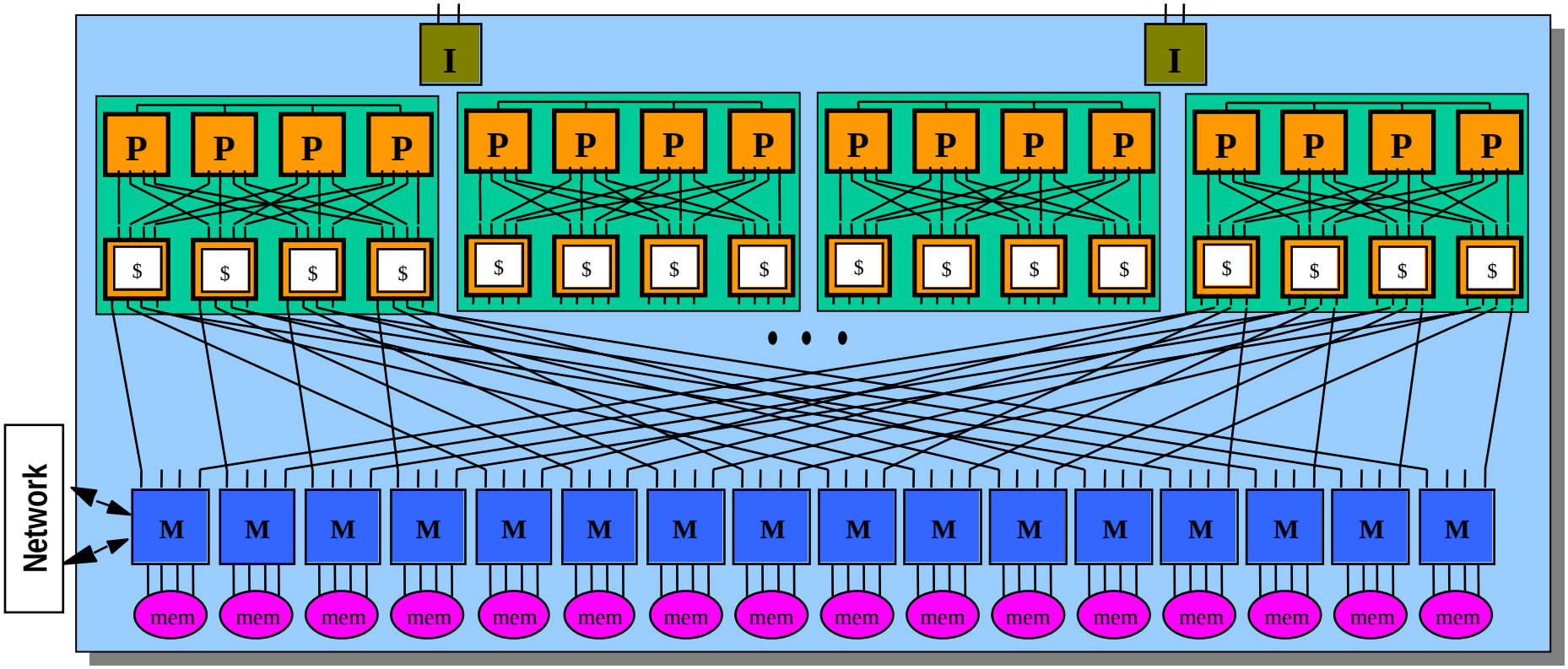
- Comprises the Forecast Model and 4DVAR
- Forecast Test Cases are
  - T21L19
  - T21L31
  - T_L159L60
  - T_L399L62
  - T_L511L60
  - T_L799L91
- Forecast Model is well understood

Description of RAPS8 (cont)

- **4DVAR Test Cases**
  - **T_L159L60**
  - **T_L511L60**
  - **T_L799L91**
- **New and Different Code since last Benchmark**
- **Preliminary Results**
  - **2.5 Days Work at INM**
  - **3-4 More Days Solving Problems**
- **Early Indicative Results for the Cray X1**

Cray X1 Node

Two SPC I/O channels per I-chip



Inter-node network
- Two ports per M-chip
- 1.6 GB/s peak both directions per port

Local node memory
- Peak BW = 16 slices x 12.8 GB/s/slice = 204.8 GB/s
- Capacity = 16 to 64 GB

Cray X1 and HPCD Metrics

Characteristic	Cray X1	HPCD
Peak Performance	3.2 Gflops/s in SSP Mode (12.8 in MSP Mode)	7.6 Gflops/s
Peak Interconnect	~6 Gbytes/s	~2 Gbytes/s
Processor Type	Vector	Scalar

Cray X1 and HPCD were introduced several years apart

Forecast Results for T_L399L62

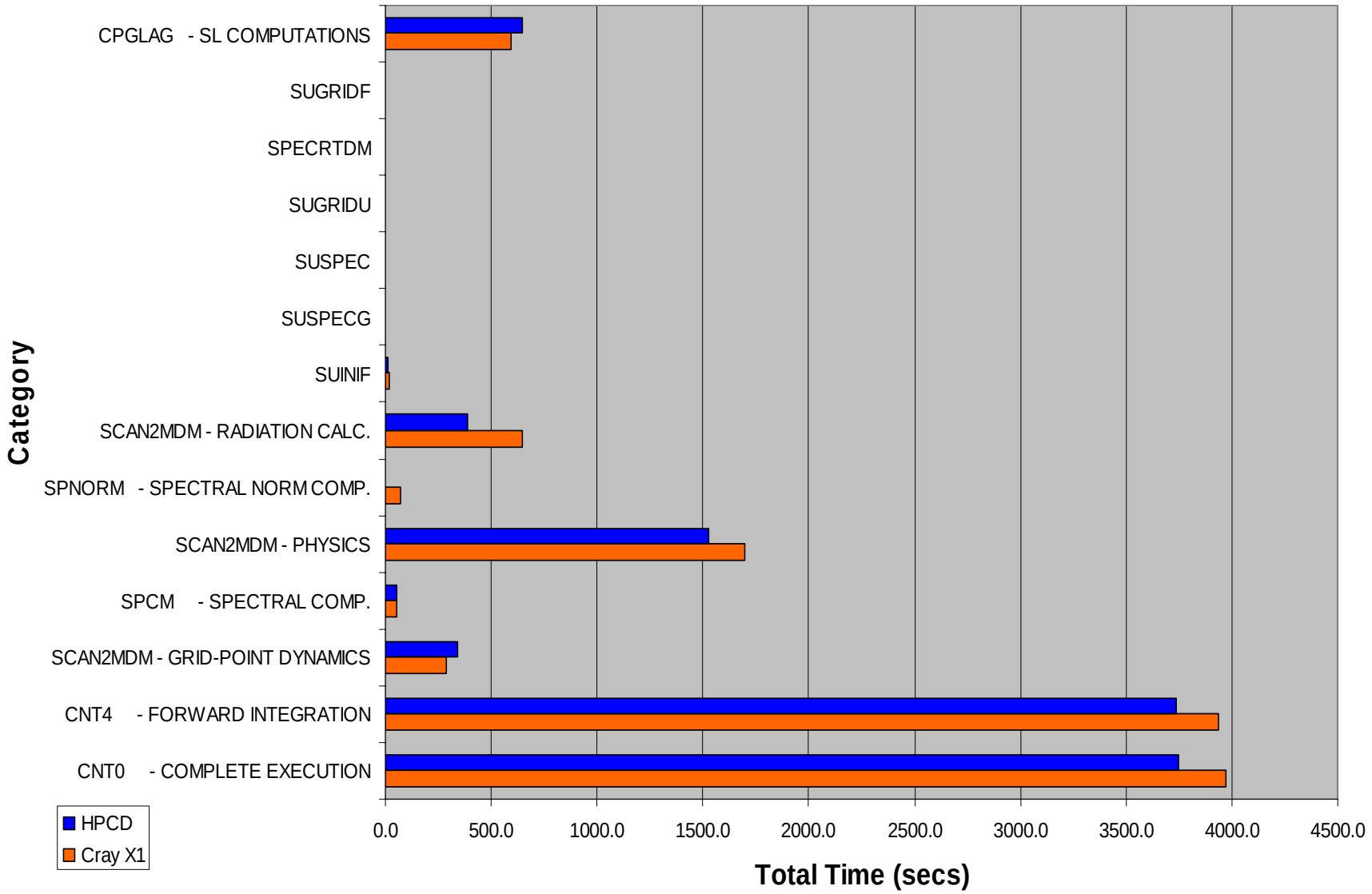
Computer	CPUs MPI x OpenMP	Wall (secs) (FCD/D)	Gflops	% of Peak
CRAY X1 at INM	64 SSPs	2102 (418)	41	20.0%
IBM p690+ at ECMWF	64 CPUs 16 x 4	2102 (412)	41	8.4%
IBM p690+ at ECMWF	128 CPUs 32 x 4	1084 (805)	80	8.2%

Forecast Results for T_L511L60

Computer	CPUs MPI x OpenMP	Wall (secs) (FCD/D)	Gflops	% of Peak
CRAY X1 at INM	120 SSPs	4042 (216)	78	20.3%
CRAY X1 at INM	240 SSPs	2026 (435)	155	20.2%
IBM p690+ at ECMWF	120 CPUs 30 x 4	4035 (215)	78	8.6%
IBM p690+ at ECMWF	240 CPUs 60 x 4	2099 (411)	150	8.2%

Total Tflop = 317

T_L511L60 GSTATS Profile



What is Dr.Hook ?

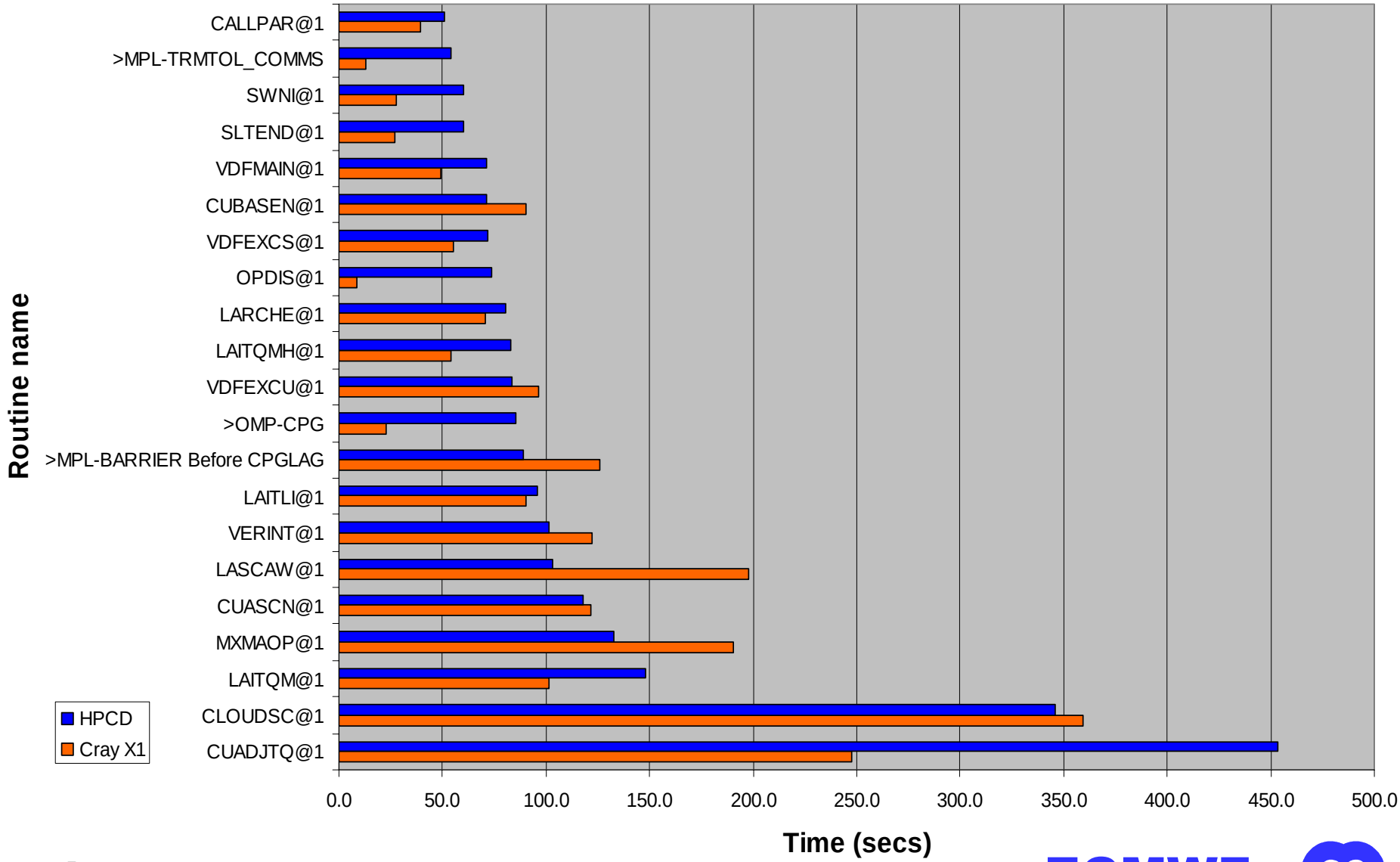
- A Fortran & C-callable instrumentation library to:
  - Trap run-time problems
  - Gather profile information per subroutine
    - Wall-clock or CPU-times
    - Mflops/s & MIPS rates
- The basic feature:
  - Keep track of the calling tree
  - Upon error (when caught via Unix-signals) tries to print the current active calling tree
  - The system specific traceback can also be printed

Dr.Hook Profiling Information

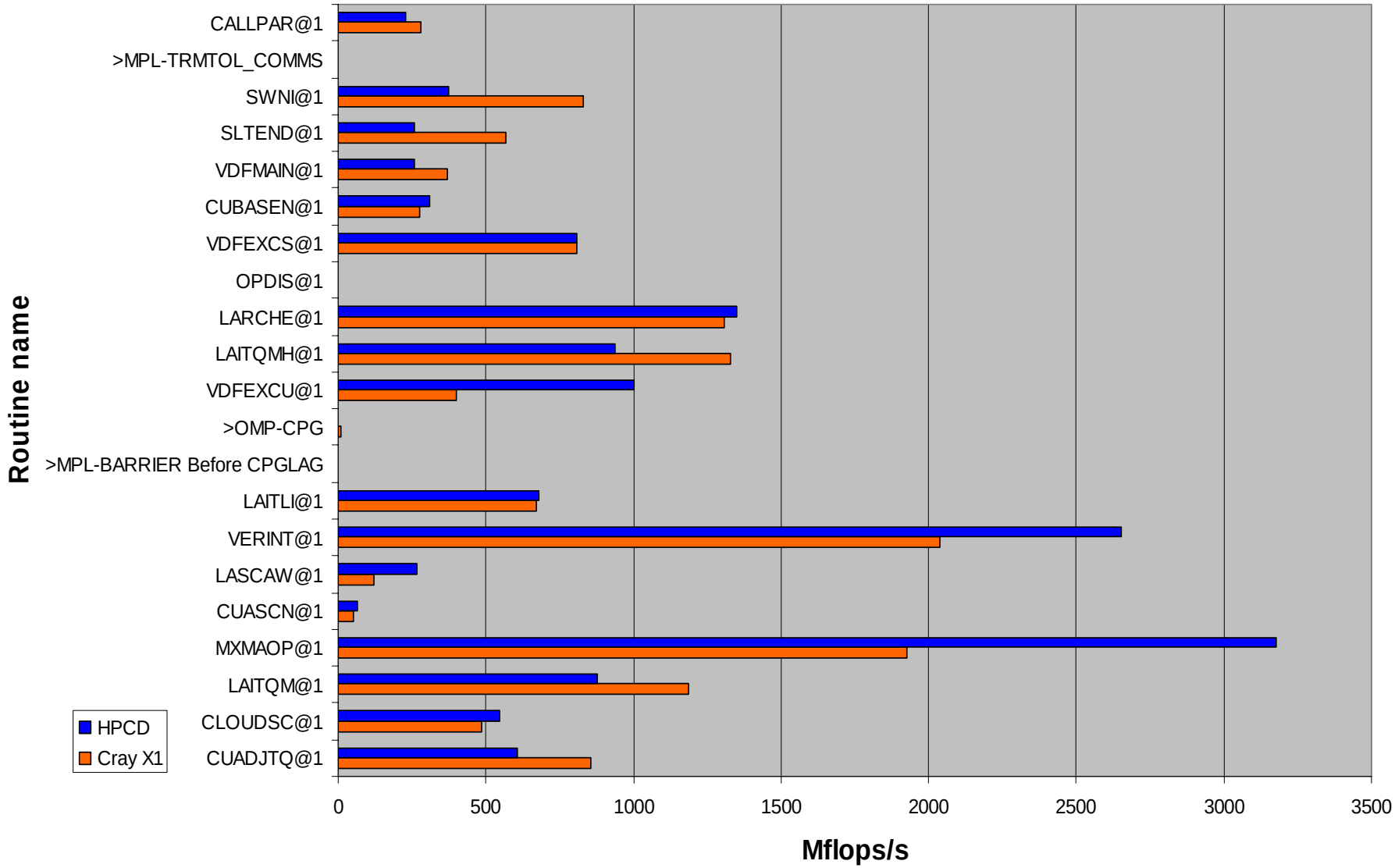
- **When the Mflops counter is enabled, the following output can be produced:**

```
Profiling information for program='/fdb/eg7t/bin/ifsMASTER', myproc#1 (# of instrumented routines called = 859):
     Instrumentation started : 20031201 171315
     Instrumentation   ended : 20031201 173631
     Wall-time is 1247.54 sec on proc#1, 401 MFlops (ops#500104*10^6), 1358 MIPS (ops#1694634*10^6) (32 procs, 4 threads)
     Thread#1:    1241.66 sec (99.53%), 124 MFlops (ops#153788*10^6), 605 MIPS (ops#751376*10^6)
     Thread#2:     505.01 sec (40.48%), 228 MFlops (ops#115265*10^6), 622 MIPS (ops#314268*10^6)
     Thread#3:     504.12 sec (40.41%), 229 MFlops (ops#115330*10^6), 626 MIPS (ops#315331*10^6)
     Thread#4:     502.39 sec (40.27%), 230 MFlops (ops#115722*10^6), 624 MIPS (ops#313659*10^6)

 #  % Time      Cumul      Self      Total    # of calls    MIPS  MFlops   Div-%   Routine@<tid> [Cluster:(id,size)]
    (self)      (sec)      (sec)      (sec)

 1   10.23    127.564    127.564    170.783       8930     685     49     0.0  *CTXGETDB@1 [57,4]
 2    5.35    194.311     66.747     98.825    7257296     807    251     0.2   *VEXP_@2 [843,4]
 3    5.35    194.311     66.688     99.131    7290992     819    255     0.2    VEXP_@4 [843,4]
 4    5.34    194.311     66.614     98.761    7298576     812    252     0.2    VEXP_@1 [843,4]
 5    5.33    194.311     66.477     98.596    7295024     808    251     0.2    VEXP_@3 [843,4]
 6    4.81    254.324     60.013    116.628    2773222     643    307     5.6  *CUADJTQ@2 [60,4]
```

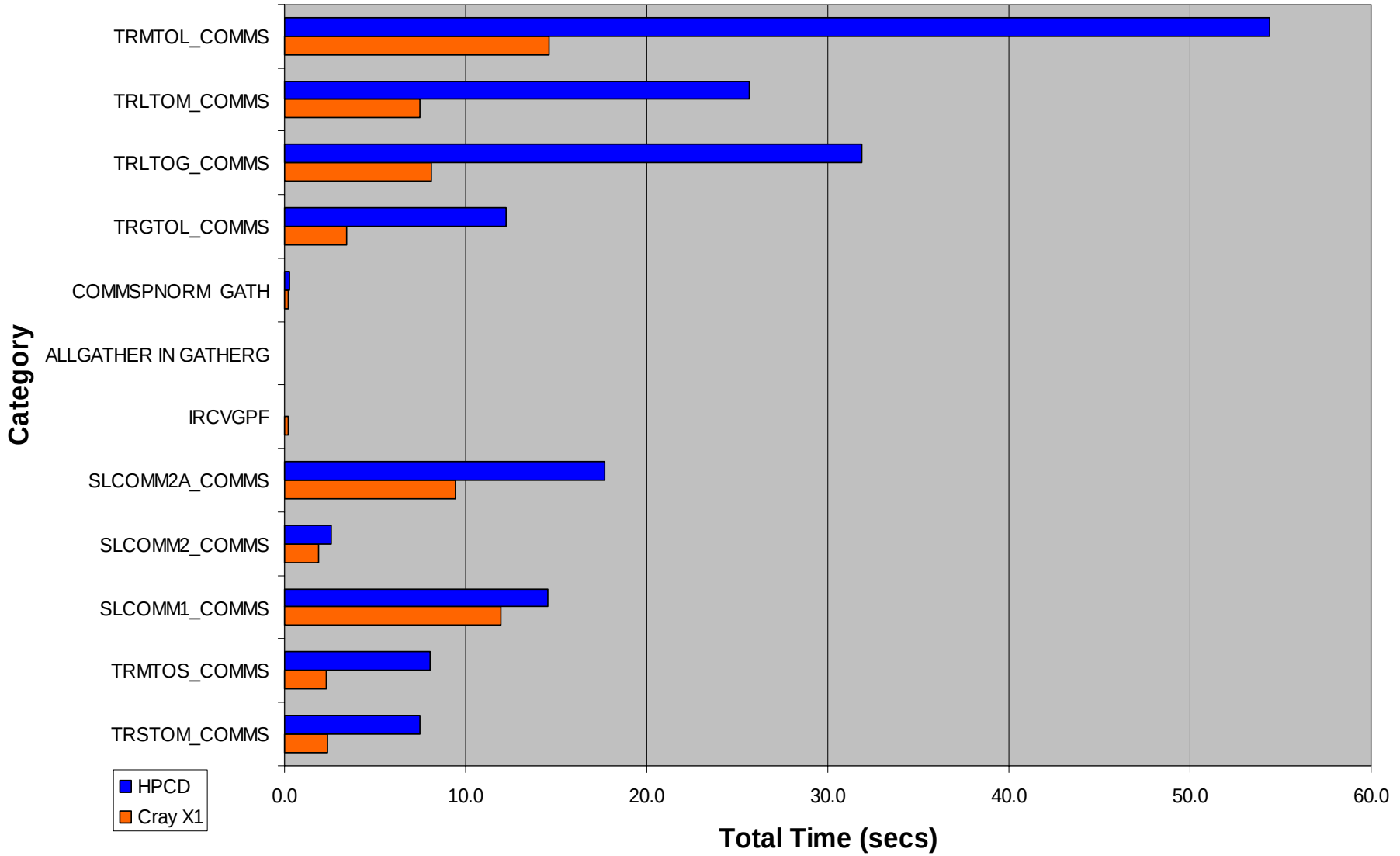

T_L511L60 Dr Hook Profile



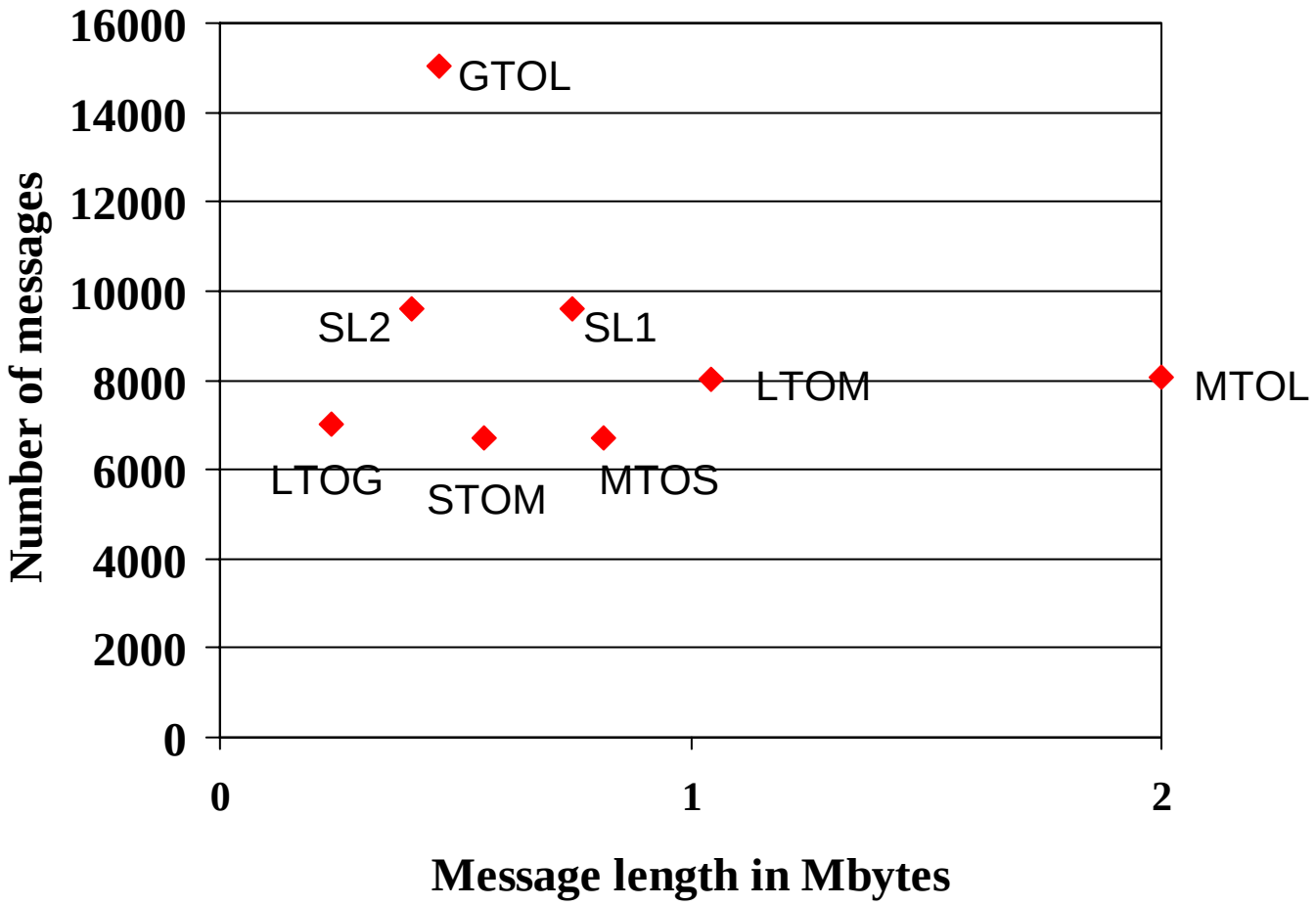
T_L511L60 Dr Hook Mflops/s



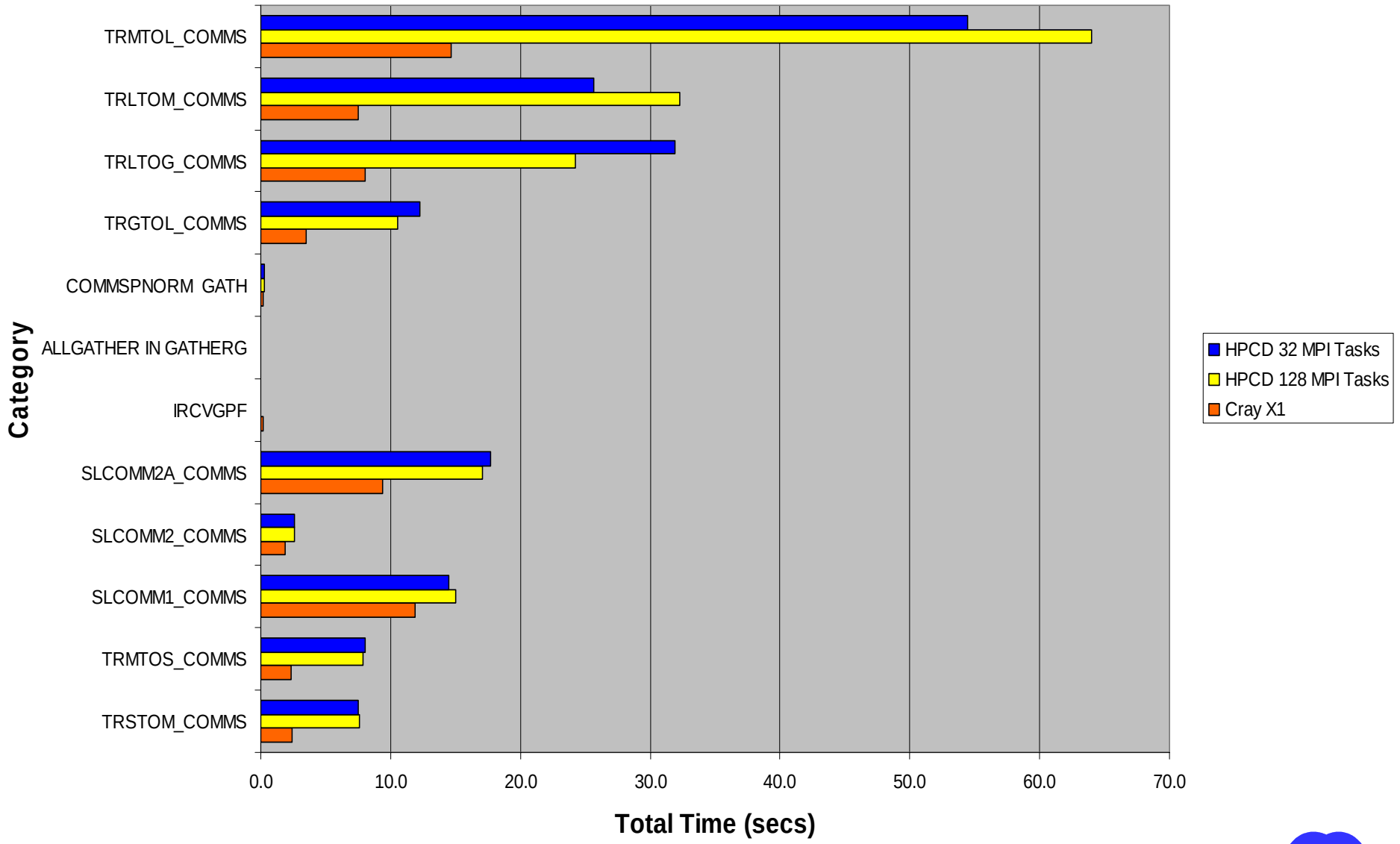
T_L511L60 32MPI GSTATS



T511 Forecast Message Lengths



T_L511L60 128/32MPI GSTATS



4DVAR

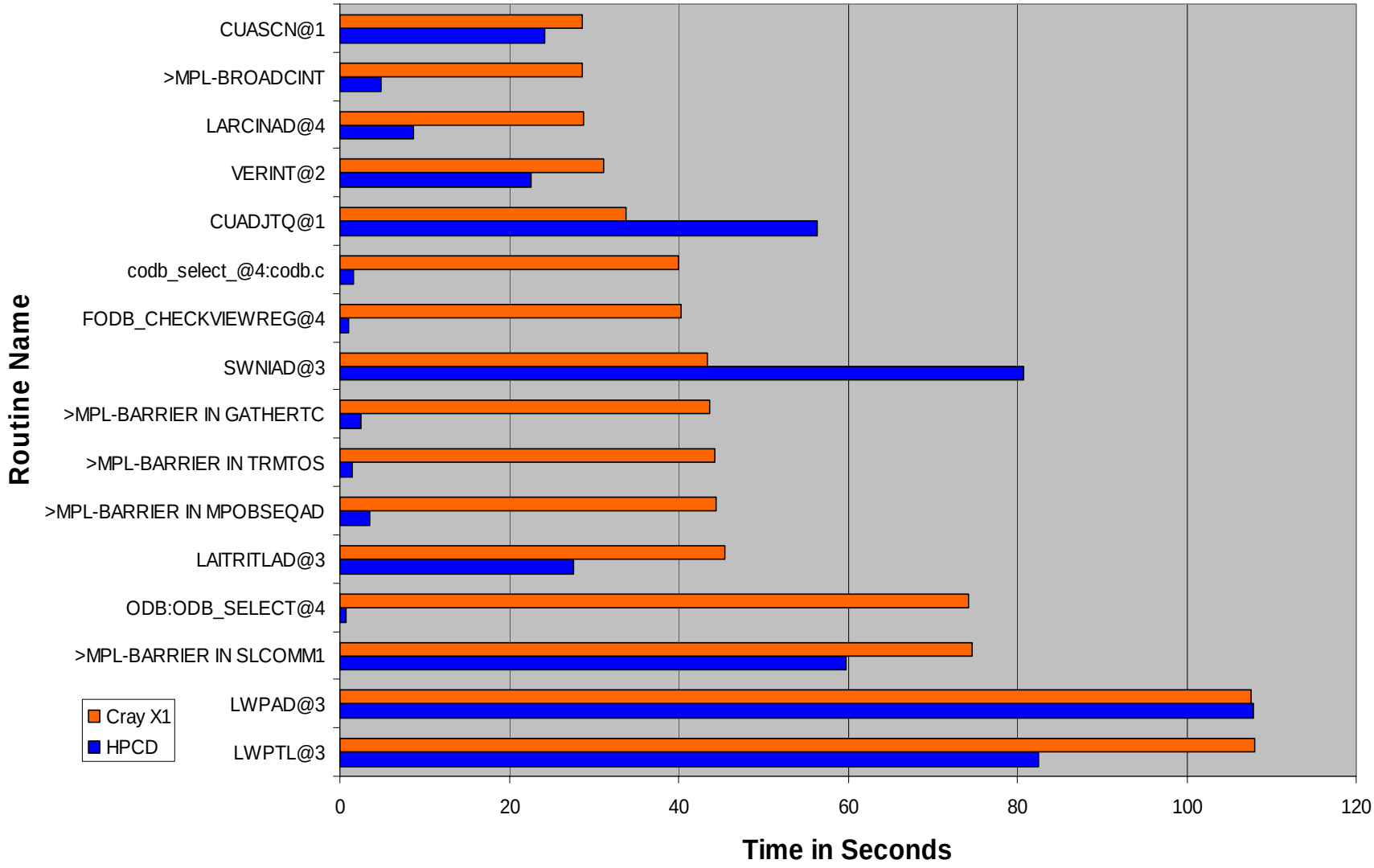
4DVAR Results for T_L511L60

Processors	Elapsed Time in Seconds Cray X1 "AS-IS"	Elapsed Time in Seconds Cray X1 after 2 Days Optimisation
128 SSP's on Cray X1	Traj_0 3524 Min_0 3728 Traj_1 434 Min_1 3150 Traj_2 1144	Traj_0 2675 Min_0 1755 Traj_1 365 Min_1 1955 Traj_2 1287

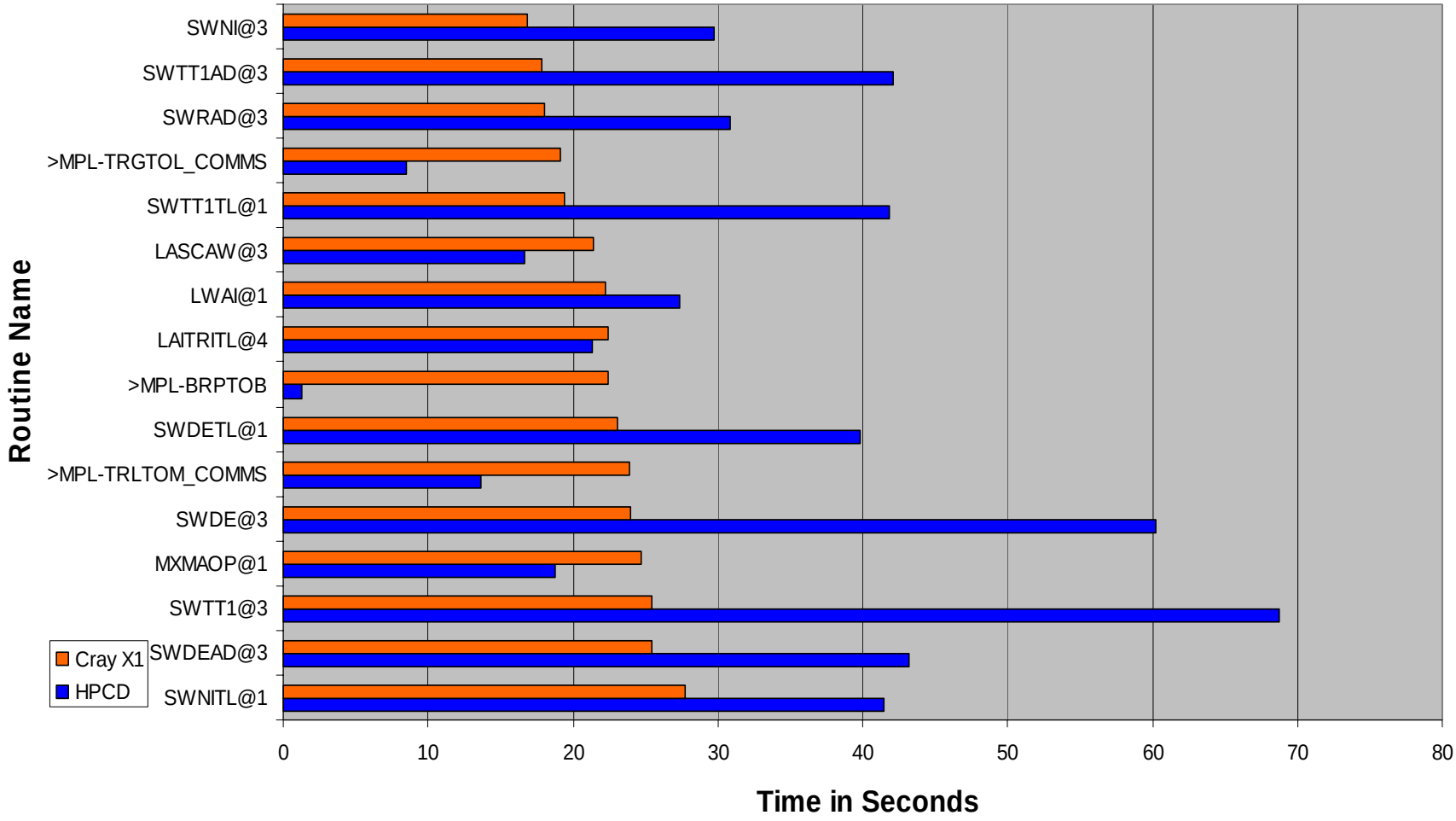
These Times are from the end of August/beginning of September

They are not the current times

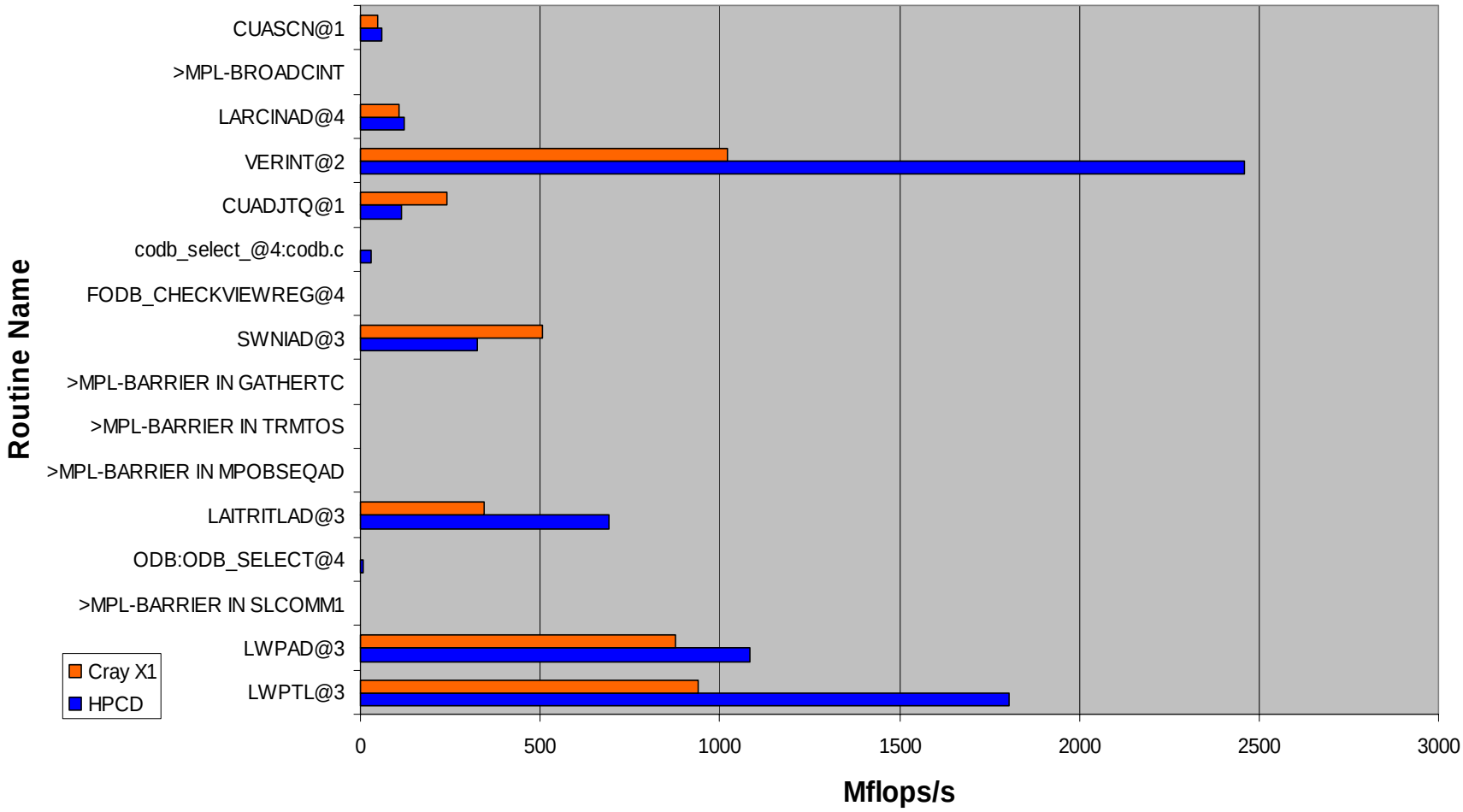
T_L511L60 MIN_1 Times - 1



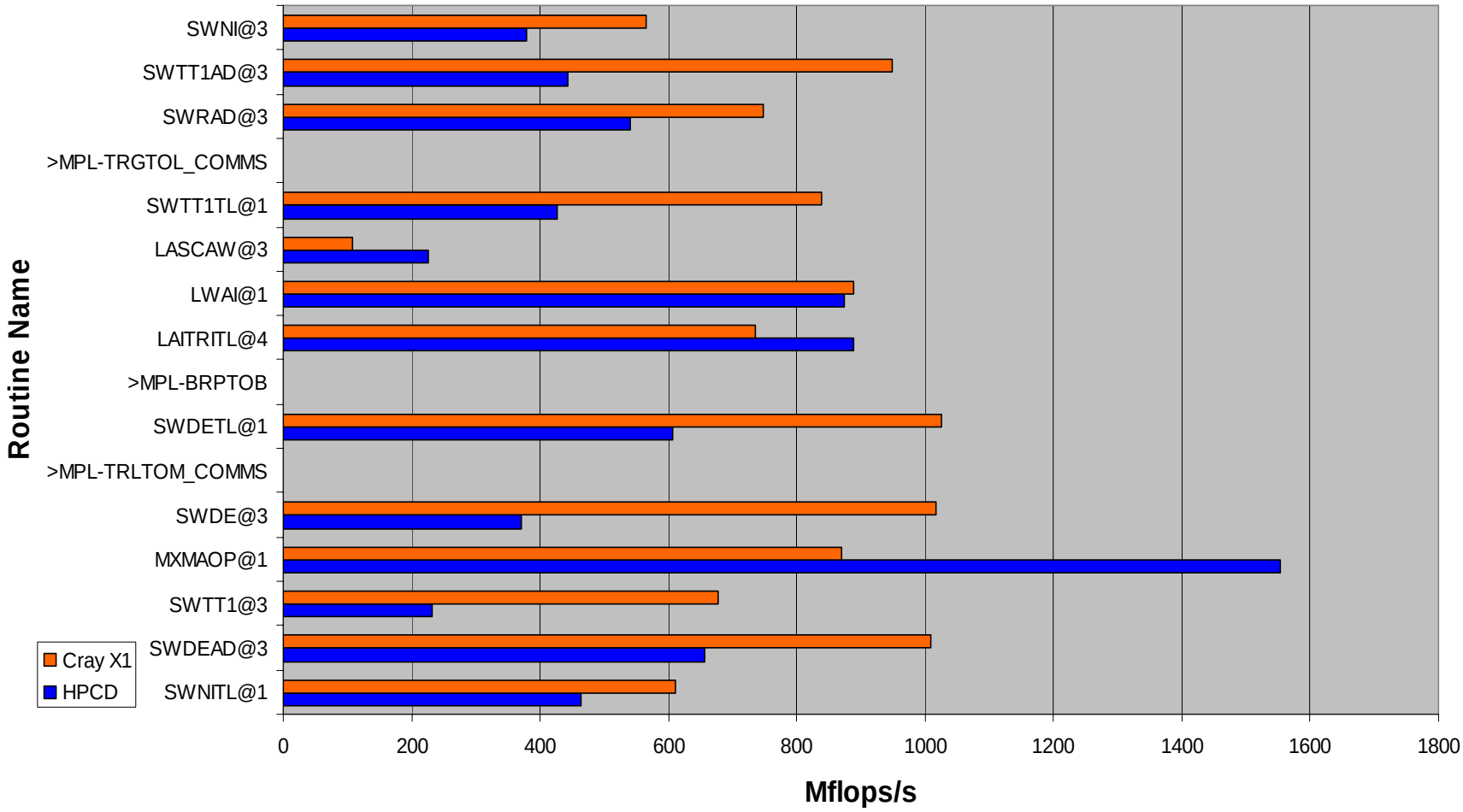
T_L511L60 MIN_1 Times - 2



T_L511L60 MIN_1 Mflops/s - 1



T_L511L60 MIN_1 Mflops/s - 2



4DVAR Optimisation

- **Early Indicative Results for the Cray X1**
  - **Further ODB Optimisation**
  - **RTTOV Optimisations**
  - **Cray Message Passing Improvements**
  - **Possibly I/O**
  - **OpenMP**
- **Plenty of Optimisation Opportunities**

Questions

Thank You

&

Any Questions?

HIRLAM 1x32 PEs

Statistics for All PEs for All Types

Total Number of Mega-Transfers	1004.764
Total Number of Mega-Words	369835.760
Total Number of Mega-Bytes	2958686.080
Total Number of Mega-Transfers per Timestep	0.930
Total Number of Mega-Words per Timestep	342.441
Total Number of Mega-Bytes per Timestep	2739.524
Total Number of Mega-Transfers per Timestep per PE	0.029
Total Number of Mega-Words per Timestep per PE	10.701
Total Number of Mega-Bytes per Timestep per PE	85.610
Average Transfer Rate	1047.933 Mbytes/s

HIRLAM 2x32 PEs

Statistics for All PEs for All Types

Total Number of Mega-Transfers	1892.329
Total Number of Mega-Words	372859.086
Total Number of Mega-Bytes	2982872.688
Total Number of Mega-Transfers per Timestep	1.752
Total Number of Mega-Words per Timestep	345.240
Total Number of Mega-Bytes per Timestep	2761.919
Total Number of Mega-Transfers per Timestep per PE	0.027
Total Number of Mega-Words per Timestep per PE	5.394
Total Number of Mega-Bytes per Timestep per PE	43.155
Average Transfer Rate	646.998 Mbytes/s