Early experiences with the new IBM p690+ at ECMWF

Deborah Salmond
& Sami Saarinen
ECMWF

Agenda

1) HPC configuration at ECMWF
   - IBM p690+

2) IFS Overview
   - MPI & OpenMP

3) Forecast on IBM p690+
   - Scalability
   - % of peak

4) 4D-Var

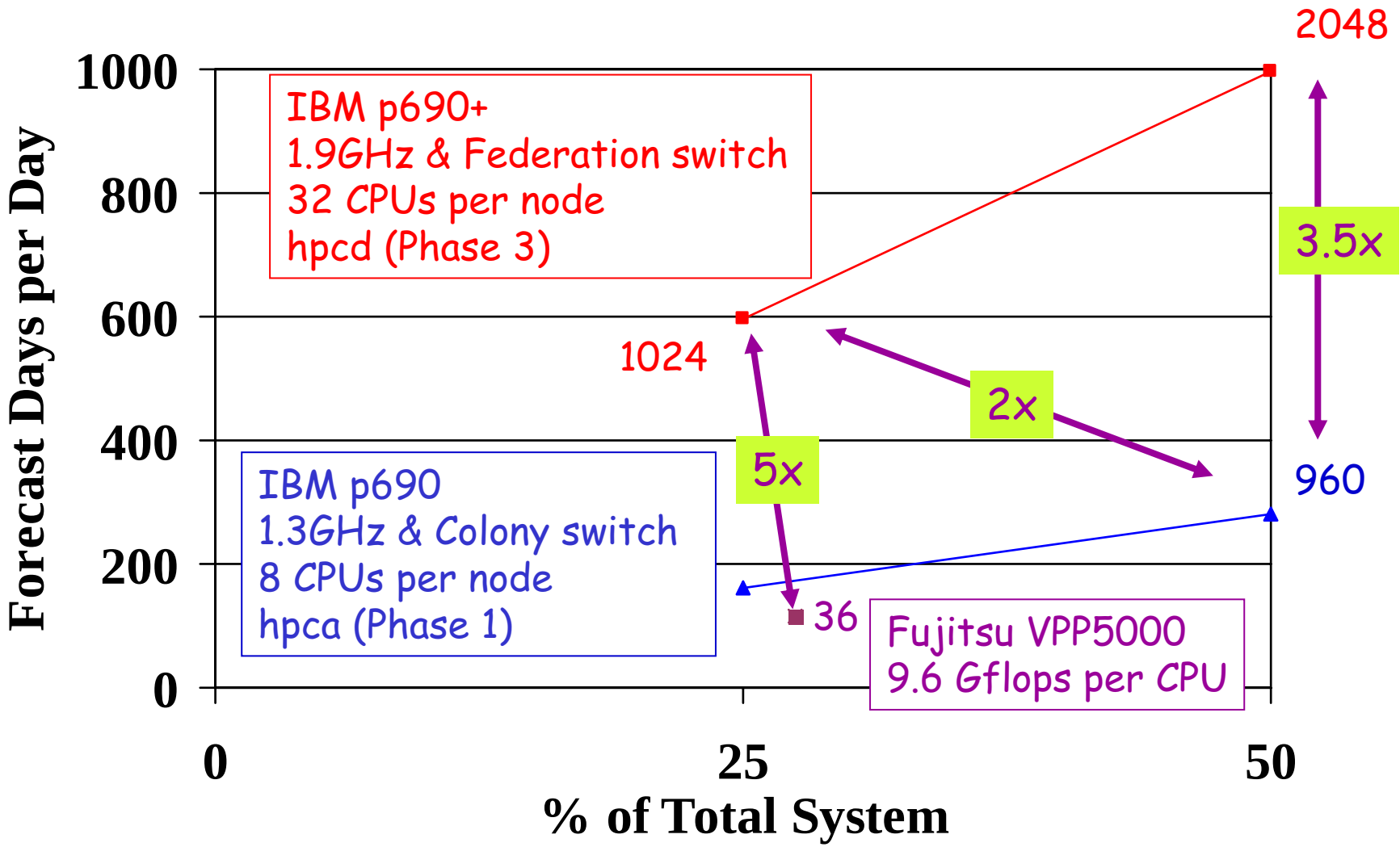
5) Optimisation and Debugging
   - Dr Hook

ECMWF - HPC configuration

Phase1 (2002): hpca & hpcb → Phase 3 (2004): hpcc & hpcd

Phase1 (2002): hpca & hpcb	Phase 3 (2004): hpcc & hpcd
IBM p690 2 x 960 processors	IBM p690+ 2 x 2176 processors
Peak performance 5.2 Gflops per processor (Power4 1.3 GHz)	Peak performance 7.6 Gflops per processor (Power4+ 1.9GHz)
Switch 350 Mbytes/s	Switch 2 Gbytes/s
8 processors per shared memory node	32 processors per shared memory node

RAPS-6 : T799 L90 Forecast benchmark



IFS - overview

-Parallelised using 'mixed' MPI and OpenMP

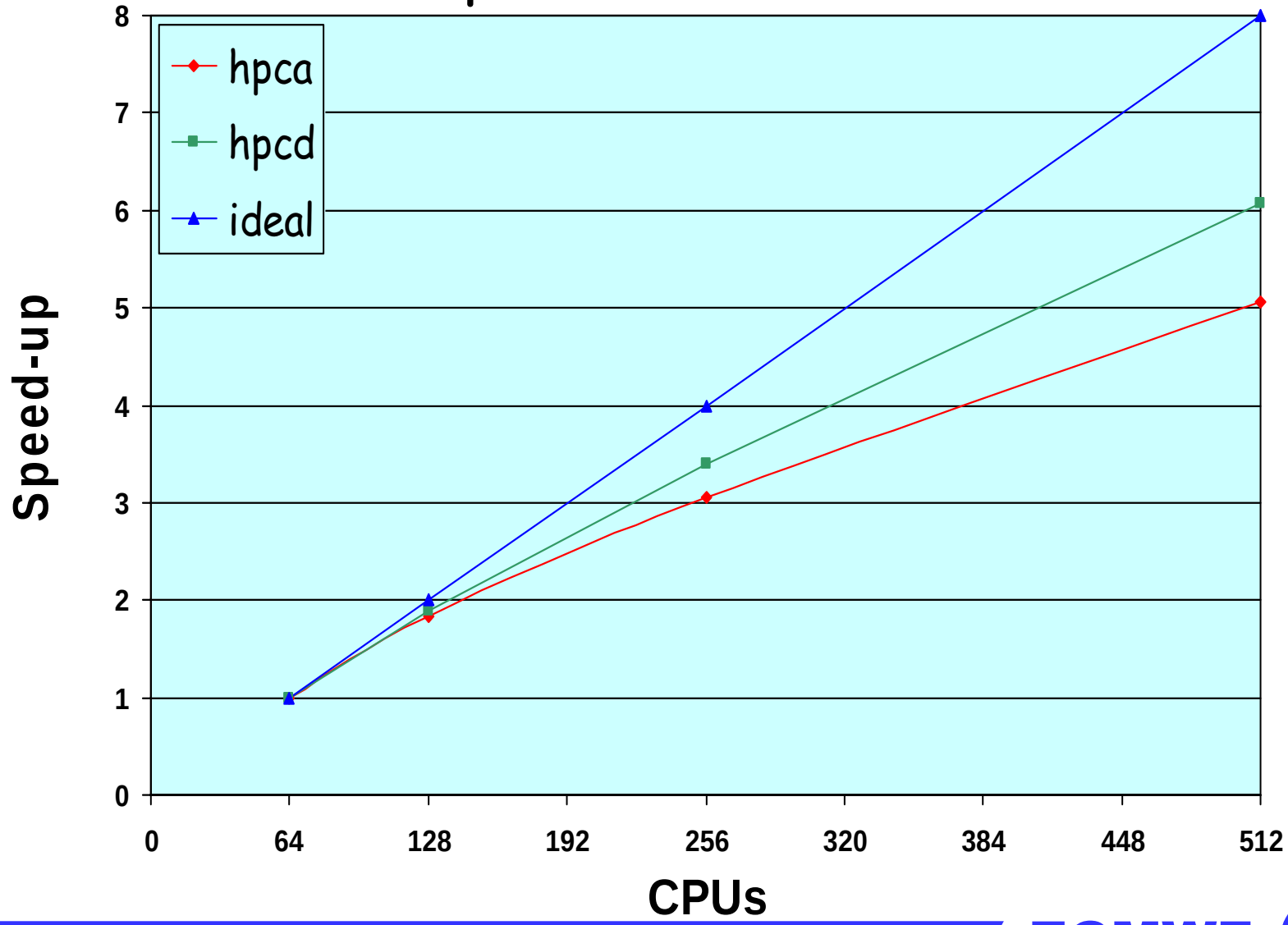
-MPI communications
- Transpositions
- Wide halo exchange
- Long messages

-OpenMP
- Shared memory nodes
- Memory efficient
- Use 4 threads

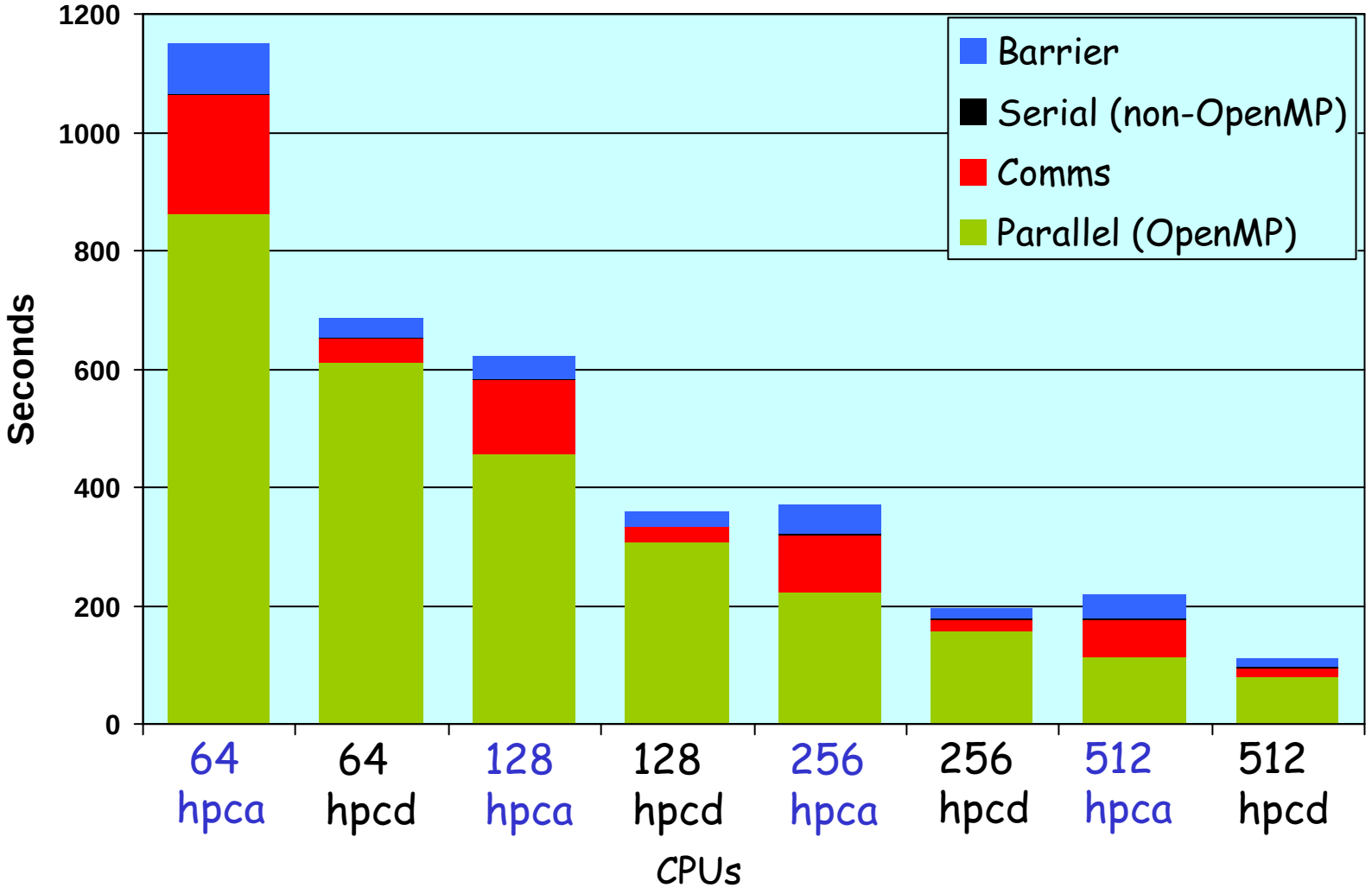
-Operational resolutions

	Current	Future
10day forecast	T511 L60 (40km)	T799 L91 (25km)
4D-Var Assimilation	T511/T95/T159 L60	T799/T95/T255 L91
EPS	T255 L40	T399 L62

T511 1-day Forecast on hpca & hpcd
4 OpenMP threads

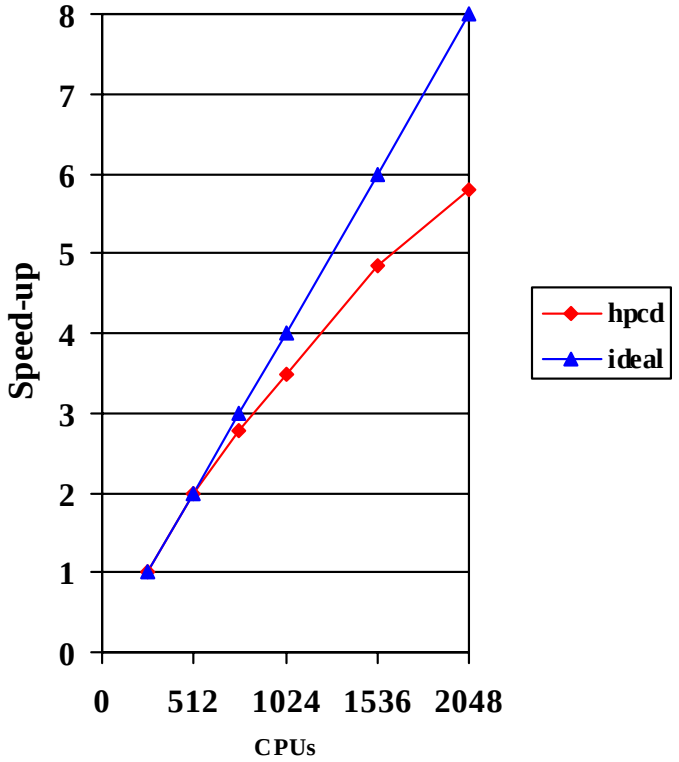


T511 1-day Forecast on hpca & hpcd
4 OpenMP threads



RAPS-8: T799 L91 10 day forecast on hpcd

MPI x OpenMP	Wall (secs)	Gflops
64x4	8850	193
128 x 4	4410	369
192 x 4	3187	509
256 x 4	2534	644
384 x 4	1830	886
256 x 8	1523	1073



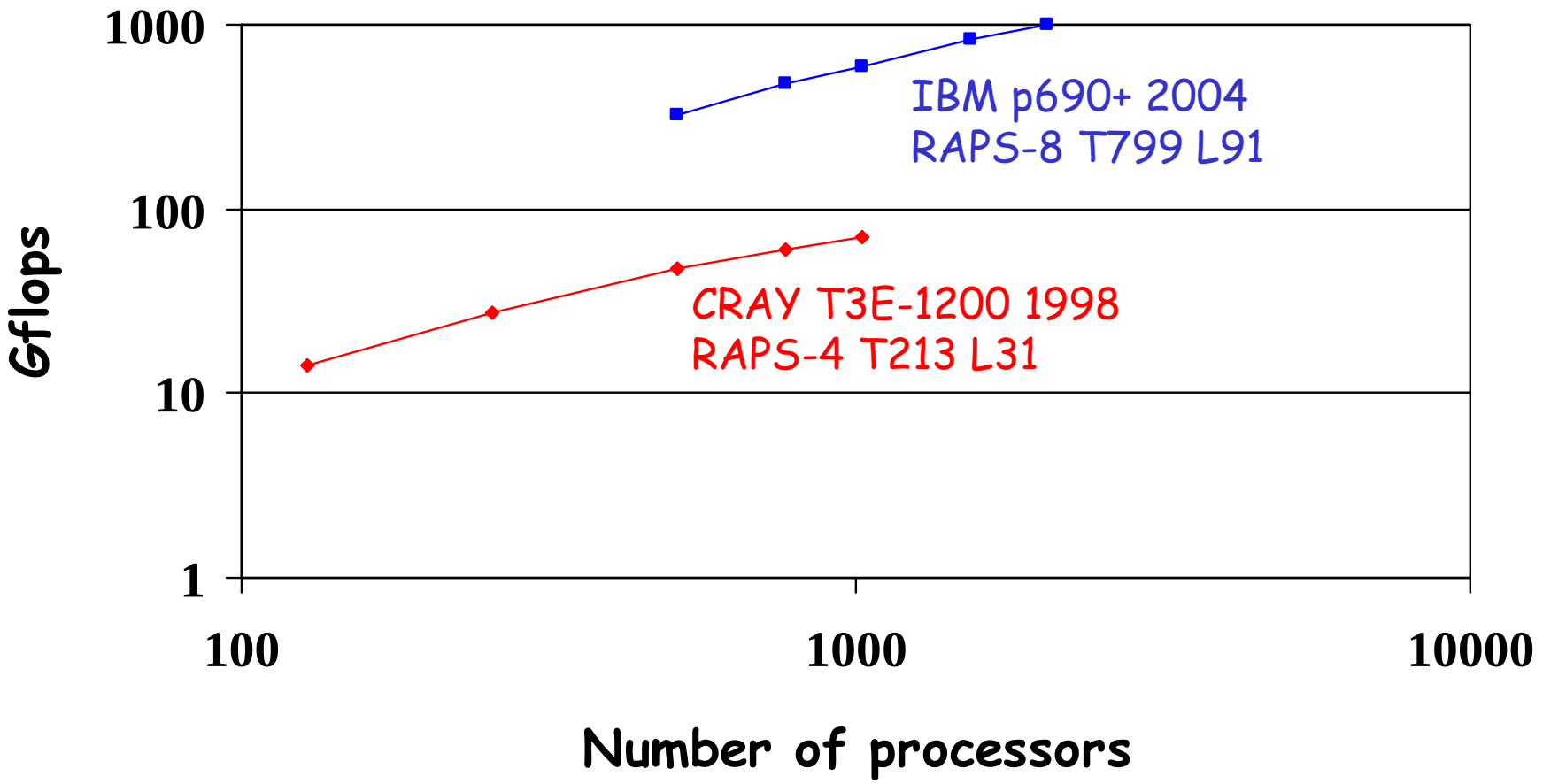
Total Pflop = 1.6

RAPS-8: T799 L91 10 day forecast on hpcd

MPI x OpenMP	Wall (secs)	Gflops	% of peak	% comms
64x4	8850	193	10.0%	5.2%
128 x 4	4410	369	9.5%	5.8%
192 x 4	3187	509	8.7%	8.3%
256 x 4	2534	644	8.2%	9.0%
384 x 4	1830	886	7.6%	10.5%
256 x 8	1523	1073	6.9%	13.2%

Total Pflop = 1.6

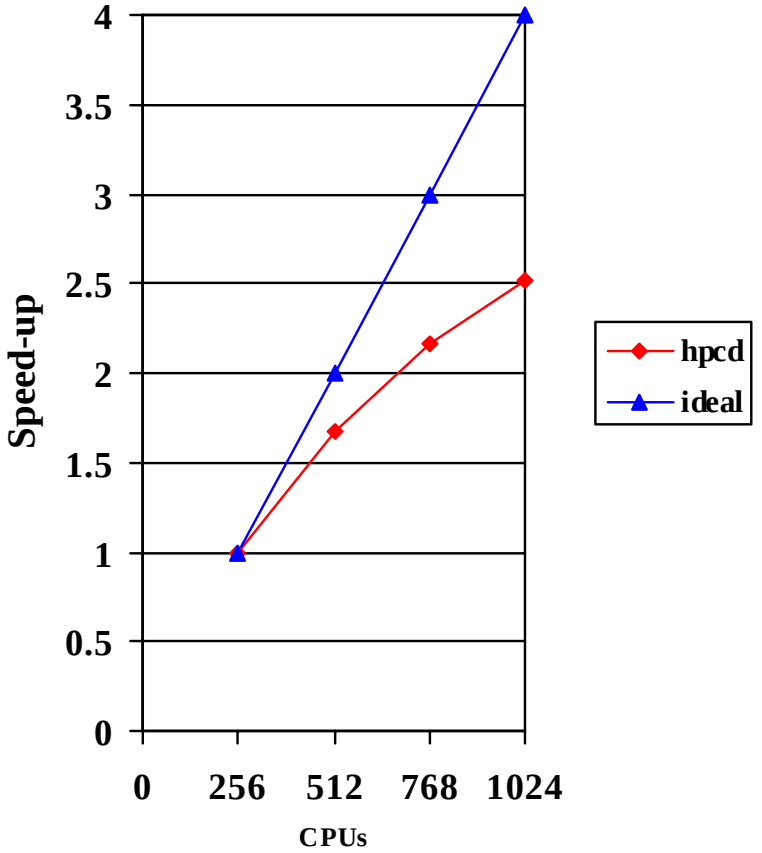
Performance of IFS Forecast on many processors



RAPS-8: 4D-Var run on hpcd

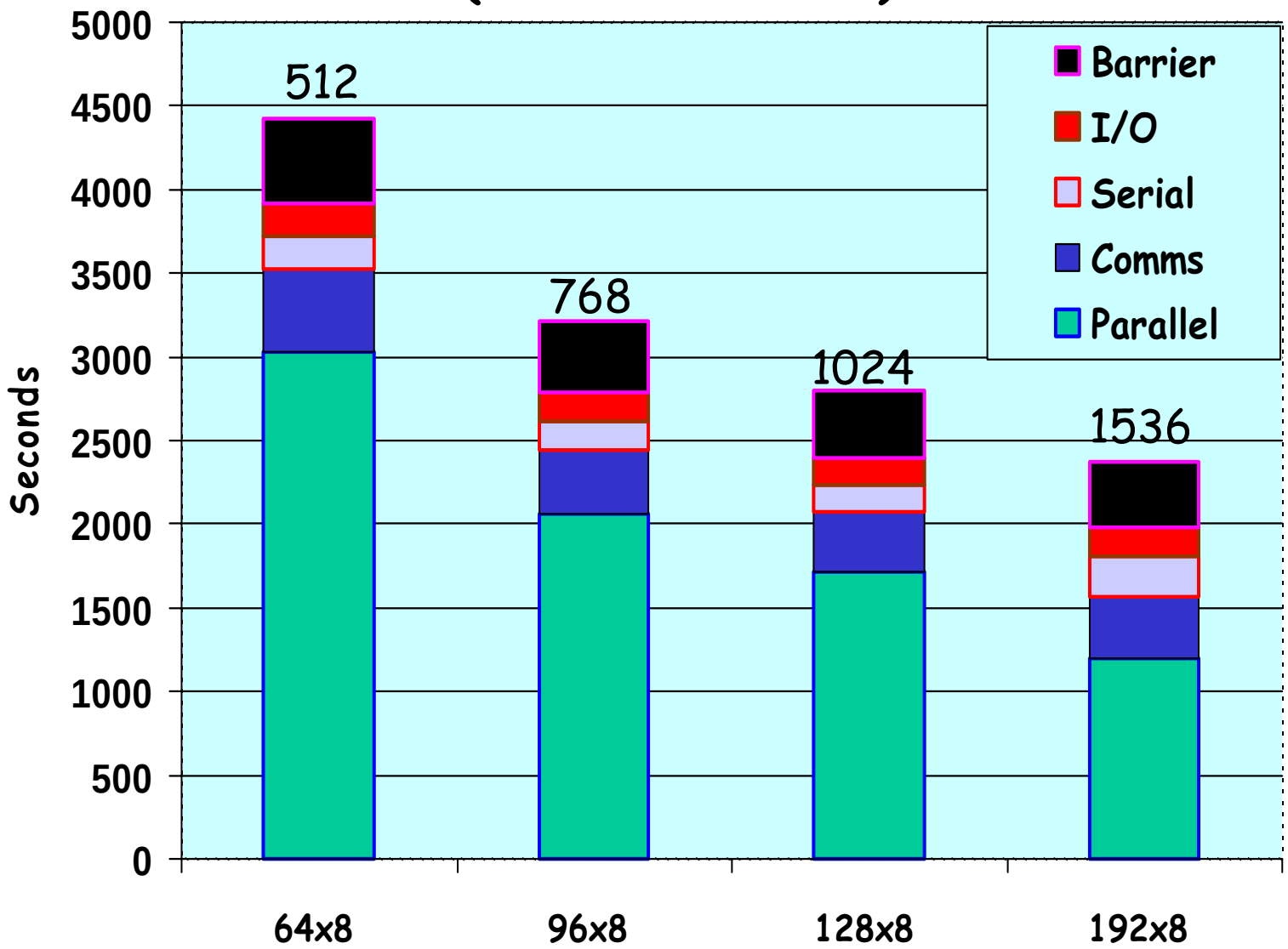
T799 L91/T95/T255

MPIx OpenMP	TOTAL (secs)	% of peak
64 x 4	5950	7.9%
128 x 4	3547	6.7%
96 x 8	2738	5.7%
128 x 8	2359	5.0%

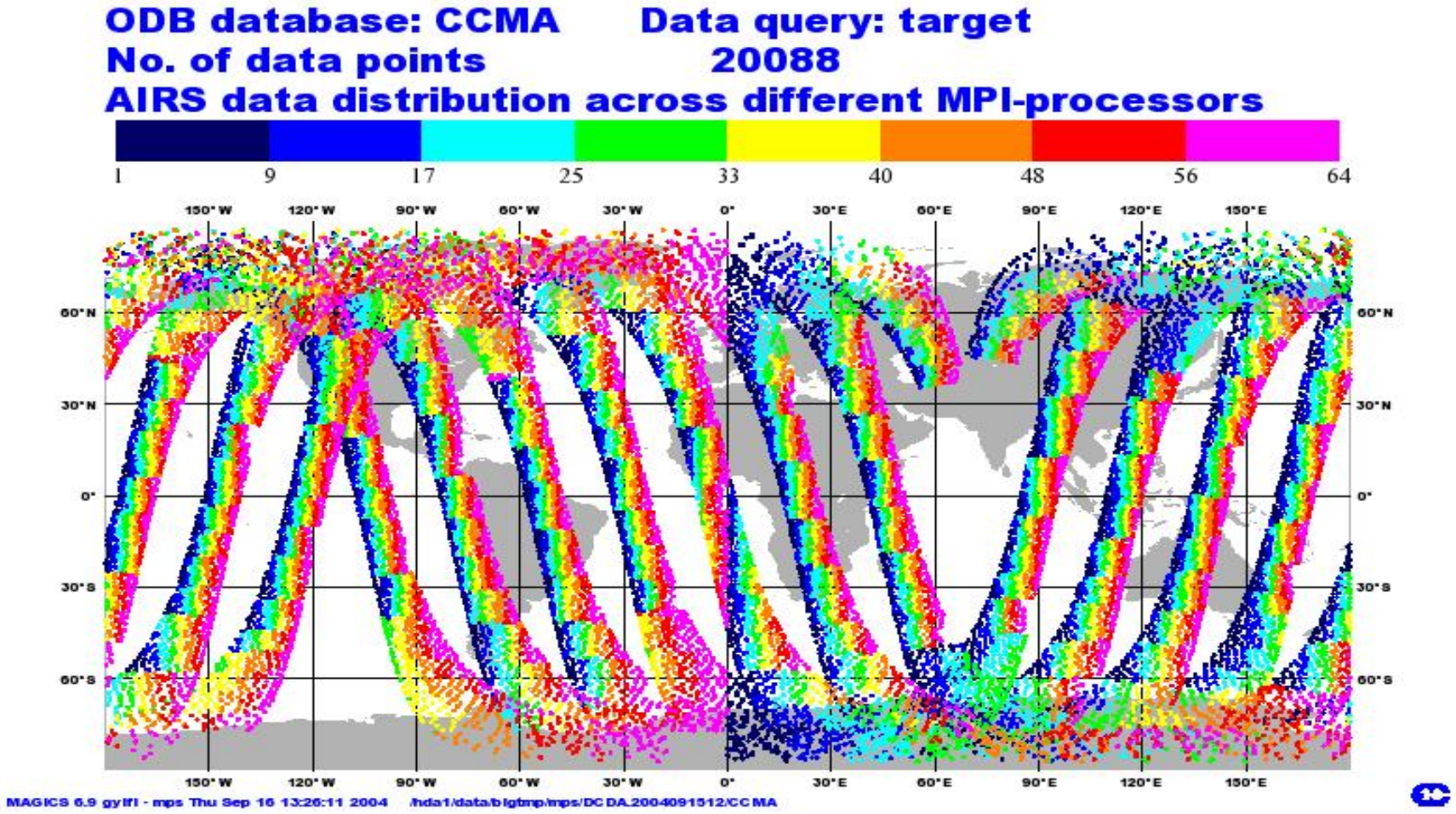


Total Tflop = 900

4D-Var T799/T95/T255 L91 on hpcd (with barriers)



Observation load-imbalance



ECMWF

Dr. Hook
- a new profiling tool

What is Dr.Hook ?

- An instrumentation library to
  - Catch run-time errors
  - Gather profile information for each instrumented subroutine :
    - Wall-clock or CPU-times
    - Mflops & MIPS – rates (IBM, Cray X1)
    - Memory usage (IBM)
- Fortran90 and C-callable

What is Dr.Hook ? *(cont'd)*

- The basic feature:
  - Keep track of the current calling tree
    - per MPI-task and OpenMP-thread
    - Upon error tries to print the calltree
  - System's own traceback can also be printed (possibly with line numbers)

- Portable with low overhead (~1% on IBM)
  - Can compare different systems

Dr. Hook Traceback

```
0:  15:57:40 STEP  936 H= 234:00 +CPU= 41.379
 13:[myproc#14,tid#4,pid#55924]: Received signal#24 (SIGXCPU) ; Memory: 2019178K (heap)
 13:[myproc#14,tid#1,pid#55924]:  MASTER ,#1,st=1,wall=0.000s/0.000s
 13:[myproc#14,tid#1,pid#55924]:   CNT0 ,#1,st=1,wall=0.000s/0.000s
 13:[myproc#14,tid#1,pid#55924]:    CNT1 ,#1,st=1,wall=0.000s/0.000s
 13:[myproc#14,tid#1,pid#55924]:     CNT2 ,#1,st=1,wall=0.000s/0.000s
 13:[myproc#14,tid#1,pid#55924]:      CNT3 ,#1,st=1,wall=0.000s/0.000s
 13:[myproc#14,tid#1,pid#55924]:       CNT4 ,#1,st=1,wall=0.000s/0.000s
 13:[myproc#14,tid#1,pid#55924]:        STEPO ,#978,st=1,wall=10531.259s/0.000s
 13:[myproc#14,tid#1,pid#55924]:         SCAN2H ,#1018,st=1,wall=8913.967s/0.043s
 13:[myproc#14,tid#1,pid#55924]:          SCAN2MDM ,#1018,st=1,wall=8913.896s/32.036s
 13:[myproc#14,tid#1,pid#55924]:           GP_MODEL ,#938,st=1,wall=8845.641s/4.830s
 13:[myproc#14,tid#1,pid#55924]:            EC_PHYS ,#213893,st=1,wall=6144.597s/22.378s
 13:[myproc#14,tid#1,pid#55924]:             CALLPAR ,#213893,st=1,wall=5856.788s/88.130s
 13:[myproc#14,tid#1,pid#55924]:              SLTEND ,#213893,st=1,wall=662.390s/179.559s
 13:[myproc#14,tid#1,pid#55924]:               CUADJTQ ,#117188599,st=1,wall=1992.364s/477.382s

 13:  Signal received: SIGXCPU - CPU time limit exceeded
 13:
 13:  Traceback:
 13:    Location 0x0000377c
 13:    Offset 0x0000009c in procedure _event_sleep
 13:    Offset 0x00000318 in procedure sigwait
 13:    Offset 0x000006c8 in procedure pm_async_thread
 13:    Offset 0x000000a4 in procedure _pthread_body
 13:    --- End of call chain ---
```

Dr. Hook profile for T511 forecast - hpca

#	%Time (self)	Cumul (sec)	Self (sec)	Total (sec)	#calls	MIPS	MFlops	%Div	Routine
1	7.43	35.027	35.027	40.573	49	961	273	2.9	WVCOUPLE@1 [567,1]
2	3.67	52.349	17.322	17.367	5824	1113	546	3.6	*CLOUDSC@1 [5,4]
3	3.65	52.349	17.204	17.287	5791	1116	548	3.6	CLOUDSC@4 [5,4]
4	3.64	52.349	17.181	17.289	5769	1118	549	3.6	CLOUDSC@2 [5,4]
5	3.63	52.349	17.138	17.202	5770	1117	549	3.6	CLOUDSC@3 [5,4]
6	3.51	68.918	16.569	16.584	54	783	0	27.6	TRMTOL_COMMS@1 [525,1]
7	2.76	81.935	13.017	18.260	51	926	1	2.8	TRGTOL@1 [520,1]
8	2.51	93.763	11.829	11.831	54	742	0	24.8	TRLTOG_COMMS@1 [523,1]
9	2.41	105.145	11.382	30.536	11540	1106	88	3.4	*CUASCN@3 [30,4]
10	2.40	105.145	11.336	30.436	11538	1112	88	3.4	CUASCN@2 [30,4]
11	2.39	105.145	11.274	30.394	11582	1110	88	3.4	CUASCN@4 [30,4]
12	2.39	105.145	11.267	30.072	11648	1113	86	3.4	CUASCN@1 [30,4]
13	2.36	116.296	11.150	11.185	3492	2135	2172	0.0	*MXMAOP@1 [166,4]
14	2.31	116.296	10.897	10.940	3502	2218	2259	0.0	MXMAOP@2 [166,4]
15	2.30	116.296	10.832	10.920	3474	2216	2258	0.0	MXMAOP@4 [166,4]
16	2.29	116.296	10.816	10.910	3484	2224	2266	0.0	MXMAOP@3 [166,4]
17	1.94	125.448	9.152	9.327	27785	1433	682	0.0	*LAITQM@3 [138,4]
18	1.94	125.448	9.130	9.263	27980	1434	679	0.0	LAITQM@1 [138,4]
19	1.92	125.448	9.073	9.256	27715	1432	682	0.0	LAITQM@4 [138,4]
20	1.92	125.448	9.045	9.220	27750	1440	686	0.0	LAITQM@2 [138,4]
21	1.85	134.173	8.725	8.785	5563	985	592	2.2	*SLTEND@4 [297,4]
22	1.85	134.173	8.724	8.777	5596	987	593	2.2	SLTEND@1 [297,4]
23	1.83	134.173	8.654	8.741	5541	986	593	2.2	SLTEND@2 [297,4]
24	1.83	134.173	8.621	8.658	5546	989	595	2.2	SLTEND@3 [297,4]
25	1.82	142.737	8.565	8.580	51	782	0	21.6	TRLTOM_COMMS@1 [524,1]
26	1.80	151.219	8.482	69.102	13	581	22	10.6	RADINTG@1 [207,1]

Dr. Hook Mflops profile for top routines in T511 forecast -> comparison of different systems

IBM p690+ (7.6 Gflops peak) MFlops per CPU	Div-%	Routine	CRAY X1 (3.2 Gflops peak) Mflops per CPU
609	2.8	CUADJTQ	857
548	4.4	CLOUDSC	486
872	0.1	LAITQM	1185
3109	0.0	MXMAOP	1925
266	0.0	LASCAW	120
1003	2.2	VDFEXCU	399
676		TOTAL	704

9%

22%

Dr. Hook memory profile for T799 forecast

```
Memory-profiling information for program='./MASTER', proc#1:
        No. of instrumented routines called : 576
Memory usage : 1505 MBytes (max.seen), 294 MBytes (leaked),
               2062 MBytes (heap), 1834 MBytes (max.rss),
               128 MBytes (max.stack), 1145 (paging)
```

#	Memory-% (self)	Self-alloc (bytes)	+ Children (bytes)	Self-Leak (bytes)	Heap (bytes)	Max.Stack (bytes)	Paging (delta)	#Calls	#Allocs	#Frees	Routine
1	20.00	587569264	554450336	76982472	2162753248	40784	51	120	723	720	GP_MODEL@1
2	9.95	292161088	78809176	0	2162753248	4404832	0	24	268	268	RADINTG@1
3	8.02	235681728	79457488	246216	2162753248	134632240	50	120	1920	1917	CALL_SL@1
4	6.52	191547280	0	0	2162753248	10530944	4	123	244	244	>TRS-FTINV
5	6.25	183480072	370970264	0	2162753248	3900704	1	24	528	528	RADDRV@1
6	4.92	144638640	0	0	698940640	20593680	0	2	10	10	>TRS-SULEG
7	4.72	138637064	134851248	96178120	968162656	18432	0	1	34	31	SUSC2B@1
8	4.59	134851248	0	134851248	871955744	19392	0	1	2	0	GMV_SUBS
9	4.39	128847896	128505720	280	698875072	13056	0	1	50	49	SUTRANS@1
10	4.19	123052080	0	0	2162753248	9690880	0	123	242	242	>TRS-LTINV
11	2.79	82091920	0	0	2162753248	9692368	1	121	242	242	>TRS-FTDIR
12	2.71	79457488	0	0	2162753248	135193904	0	120	360	360	SLCOMM2A@1
13	2.68	78809176	0	0	1998651072	4408480	0	2	6	6	SLCOMM@1
14	2.39	70328856	0	0	2162753248	10530944	4	123	121	121	>TRS-FTINV
15	2.18	64001688	400	1664	423754272	13024	0	1	8	3	SUMPINI@1
16	1.63	47978864	13602032	0	1016200576	17115456	1	1	1	1	SUGRIDUG@1
17	1.56	45774496	144638640	280	698940640	27184	835	1	113	112	SUECRAD@1
18	1.51	44311240	0	0	2162753248	134635936	0	120	360	360	SLCOMM1@1
19	1.31	38481456	0	0	2162753248	21983104	0	120	240	240	TRSTOM@1
20	1.27	37441424	0	0	2162753248	21982880	0	120	240	240	TRMTOS@1
21	1.05	30763032	0	0	2162753248	9692368	0	121	121	121	>TRS-FTDIR
22	1.05	30763032	82091920	0	2162753248	672608	0	120	240	240	TRANSDIR_MDL@1
23	0.79	23068936	0	40	698875072	13264	0	1	4	3	SUALSPA@1
24	0.70	20555752	0	0	2162753248	4701600	0	46	138	138	SLCOMM2@1

Dr. Hook memory profile for T799 forecast

#	Memory-% (self)	Self-alloc (bytes)	+ Children (bytes)	Routine
1	20.00	587569264	554450336	GP_MODEL@1
2	9.95	292161088	78809176	RADINTG@1
3	8.02	235681728	79457488	CALL_SL@1
4	6.52	191547280	0	>TRS-FTINV
5	6.25	183480072	370970264	RADDRV@1
6	4.92	144638640	0	>TRS-SULEG
7	4.72	138637064	134851248	SUSC2B@1
8	4.59	134851248	0	GMV_SUBS
9	4.39	128847896	128505720	SUTRANS@1
10	4.19	123052080	0	>TRS-LTINV

Dr. Hook memory profile for T799 forecast

#Calls	#Allocs	#Frees	Routine
120	723	720	GP_MODEL@1
24	268	268	RADINTG@1
120	1920	1917	CALL_SL@1
123	244	244	>TRS-FTINV
24	528	528	RADDRV@1
2	10	10	>TRS-SULEG
1	34	31	SUSC2B@1
1	2	0	GMV_SUBS
1	50	49	SUTRANS@1
123	242	242	>TRS-LTINV

Top Optimisation activities for IFS

1. Add timings
2. Improve MPI comms (not buffered and no overlap)
3. Add more OpenMP parallel regions
4. Divides (-qstrict)
5. Use vector functions and machine specific libraries
6. Remove copies and zeroing of arrays
7. Optimise data access
8. Remove low level allocates