

## CASTEP Benchmarking Results

Benchmark	Nprocs	XT (bespoke FFTs)	X2 (bespoke FFTs)	X2 (libsci FFTs)	Speedup factor XT:X2bespoke FFTs
allx1 (step 7)	1	2042.44s	1464.70s	1446.66s	1.39
	8	437.81s	243.85s	307.55s	1.80
	<b>16</b>	<b>231.92s</b>	<b>157.28s</b>	<b>188.83s</b>	<b>1.47</b>
	32	166.61s	128.62s	143.3s	1.30
al3x3 (step 11)	<b>16</b>	<b>10152s*</b>	<b>4436s</b>	<b>4648s</b>	<b>2.29</b>
	32	6077.90s	2162.73s	2263.76s	2.81
	64	3352.89s	1225.48s	1274.03s	2.73
	112	2196.7s	912.55s	930.58s	2.41
TiN-mp (step 39)	16	3327.63s	1651.2	2107.37s	2.02
	32	1629.41s	1256.22s	1454.51s	1.30
	64	832.52s	2066.07s	2227.79s	**

- Timings taken from CASTEP output on the SCF step indicated, except al3x3, 16 proc. The timings for this job were taken from aprun's wall clock time in the batch output file: CASTEP's internal timer returned the wrong values (around 4 times slower than actual).
- \*This job was run in single core mode due to memory requirements.
- \*\*The anomalous result for TiN-mp on 64 processes produced, for each run, the following message at the end of the SCF cycles: “\*Warning\* max. SCF cycles performed but system has not reached the groundstate”
- CASTEP's own FFTs are consistently faster than Cray's libsci FFTs for these tests.
- CASTEP FFTs timed here used the values lotmax=128, lvr=128 (the length of a vector register) and the IVDEP (ignore vector dependencies) directive was used on the key FFT loops. Performance doesn't seem to be too sensitive to different values of lotmax, unless it's set to an inappropriate value such as 1 (see below).
- Cases in bold have been investigated with CrayPAT; see below.

### Effect of lotmax

The performance effect of varying the value of the parameter lotmax (a blocking parameter for multiple 1-d FFTs) in the bespoke FFTs for 16-processor allx1 jobs on the X2 (timing on the TDS machine):

lotmax=1	lotmax=64	lotmax=128	lotmax=256	lotmax=512
300.99s	175.92	175.66	175.64	175.96

128 = number of elements in a vector register. Beyond a value of 1, this doesn't appear to have an effect on the performance of CASTEP.

## CrayPAT comparisons

The command `pat_report- Oapa` builds a sampling experiment that is used to generate a more informative tracing experiment (see below). However, the output from the sampling experiment is also useful to get a basic profile by function showing where time is spent. The first set of results are for the `allx1` case running on 16 processors.

## CrayPAT Sampling

### XT allx1 16 proc

Samp %	Samp	Imb. Samp	Imb. Samp %	Group Function
100.0%	16913	--	--	Total
54.5%	9210	--	--	ETC
15.0%	2534	70.81	2.9%	zgemm_kernel_n
8.0%	1359	59.94	4.5%	zgemm_kernel_l
4.1%	700	46.50	6.6%	zgemm_otcopy
2.7%	462	26.94	5.9%	zcopy_k
2.6%	442	35.81	8.0%	zgemm_ncpy
2.2%	367	172.88	34.1%	PtlEQPeek
2.2%	364	35.31	9.4%	zlaset_
1.6%	267	23.56	8.6%	zlasr_
1.4%	244	16.69	6.8%	zdotu_
1.2%	205	17.75	8.5%	zlacpy_
1.0%	177	20.94	11.3%	zscal_k
1.0%	161	27.31	15.5%	zgemv_n
24.3%	4105	--	--	MPI
17.5%	2956	105.38	3.7%	mpi_alltoallv_
4.3%	725	48.06	6.6%	mpi_recv_
2.2%	365	20.62	5.7%	mpi_allreduce_
21.3%	3598	--	--	USER
5.3%	904	58.31	6.5%	COMMS_TRANSPOSE_EXCHANGE.in.COMMS_TRANSPOSE_N.in.COMMS
2.4%	402	36.44	8.9%	RAD5I.in.FFT_GPFA
2.3%	397	35.25	8.7%	GPFA.in.FFT_GPFA
1.3%	225	25.88	11.0%	RAD4ITWID.in.GPFA2F.in.FFT_GPFA
1.3%	223	36.12	14.9%	RAD4II.in.GPFA2F.in.FFT_GPFA
1.1%	188	18.50	9.6%	ION_BETA_RECIP_INTERPOLATION.in.ION

### X2 bespoke FFTs allx1 16 proc

Samp %	Samp	Imb. Samp	Imb. Samp %	Experiment=1 Group Function
100.0%	22552	--	--	Total

Samp %	Samp	Imb.	Imb.	Experiment=1
		Samp	Samp %	Group
				Function
				PE= 'HIDE'
59.4%	13386	--	--	ETC
13.0%	2933	246.88	8.3%	zgbmv_
9.8%	2213	136.12	6.2%	zgemm_
8.1%	1817	150.19	8.1%	MPIDI_CRAY_progress
3.0%	677	81.12	11.4%	MPIDI_CRAY_dmdev_progress
2.4%	530	67.69	12.1%	_F90_LEN_TRIM_
1.9%	433	266.94	40.7%	getrusage
1.9%	433	53.81	11.8%	zcopy_
1.9%	427	57.06	12.6%	zhbm_
1.4%	324	52.12	14.8%	MPIDI_CRAY_Progress_wait
1.4%	307	51.56	15.3%	_F90_FCD_CMP_EQ
1.2%	270	36.44	12.7%	MPIC_Wait
1.0%	218	25.75	11.3%	zdotu_
22.2%	5007	--	--	USER
4.1%	920	110.56	11.4%	trace_entry\$trace_
3.9%	879	125.50	13.3%	ion_beta_recip_interpolation\$ion_
2.6%	592	566.81	52.2%	comms_transpose_exchange\$comms_transpose_n\$comms_
1.9%	422	76.00	16.3%	gpfa2f\$fft_gpfa_
1.1%	258	46.50	16.3%	gpfa\$fft_gpfa_
1.1%	237	39.00	15.1%	gpfa5f\$fft_gpfa_
18.4%	4159	--	--	MPI
13.7%	3087	356.81	11.1%	mpi_alltoallv_
3.1%	696	79.69	11.0%	mpi_allreduce_
1.1%	250	37.50	13.9%	mpi_recv_

### X2 libsci FFTs allx1 16 proc

Samp %	Samp	Imb.	Imb.	Experiment=1
		Samp	Samp %	Group
				Function
				PE= 'HIDE'
100.0%	26110	--	--	Total
68.6%	17902	--	--	ETC
11.2%	2914	248.56	8.4%	zgbmv_
8.5%	2216	165.06	7.4%	zgemm_
6.6%	1718	244.31	13.3%	MPIDI_CRAY_progress
6.0%	1578	182.50	11.1%	zfftx32_
3.1%	816	124.44	14.1%	MPIDI_CRAY_dmdev_progress
2.0%	511	130.75	21.7%	_F90_LEN_TRIM_
1.8%	472	56.06	11.3%	zpassm1\$32_
1.7%	454	146.00	26.0%	getrusage
1.7%	436	36.69	8.3%	zcopy_
1.7%	435	143.75	26.5%	zpassl\$32_
1.7%	432	51.06	11.3%	zpass\$32_
1.6%	419	32.88	7.8%	zhbm_
1.2%	306	53.69	15.9%	MPIDI_CRAY_Progress_wait
1.2%	302	70.19	20.1%	_F90_FCD_CMP_EQ
1.1%	279	64.31	20.0%	MPIC_Wait
1.0%	262	53.62	18.1%	zpasslf_r6\$32_
16.2%	4238	--	--	MPI

```

-----
| 12.0% | 3144 | 223.56 | 7.1% | mpi_alltoallv_
| 2.6% | 683 | 81.69 | 11.4% | mpi_allreduce_
| 1.1% | 281 | 24.56 | 8.6% | mpi_recv_
=====
| 15.2% | 3970 | -- | -- | USER
-----
| 3.5% | 908 | 115.81 | 12.1%
ion_beta_recip_interpolation$ion_
| 3.4% | 897 | 105.19 | 11.2% | trace_entry$trace_
| 2.1% | 538 | 67.31 | 11.9%
|comms_transpose_exchange$comms_transpose_n$comms_
=====

```

These tables show that most time is spent in the BLAS, so CASTEP would benefit greatly from an X2-tuned version. Two banded matrix routines appear in the X2 profiles and not in those for the XT (and are therefore insignificant on the XT): zhbmv and zgbmv. This suggests that banded matrices in general are poorly handled on the X2 with libsci BLAS. The poor performance of the NaHF2 benchmark shows a particular problem with the BLAS. Running on a single processor, the following times are produced:

XT	X2 libsci BLAS
467.2s	544.8s

Profiling shows that this slowdown is due to poor vectorization of the BLAS routine zhbmv (a complex Hermitian banded matrix-vector operation). After downloading this routine from netlib, compiling (with only `ftn -c`) and linking, the timing was reduced to 361.25s. The non-vectorized main loop is:

```

DO 60 J = 1,N
  TEMP1 = ALPHA*X(J)
  TEMP2 = ZERO
  L = KPLUS1 - J
  DO 50 I = MAX(1,J-K),J - 1
    Y(I) = Y(I) + TEMP1*A(L+I,J)
    TEMP2 = TEMP2 + DCONJG(A(L+I,J))*X(I)
50  CONTINUE
  Y(J) = Y(J) + TEMP1*DBLE(A(KPLUS1,J)) + ALPHA*TEMP2
60  CONTINUE

```

The references to Y(I) and Y(J) constitute a dependency, thus blocking vectorization of the outer loop. However, in CASTEP's call to this routine for NaHF2 the argument K (number of super-diagonals) is zero, which means the inner loop is never exercised. Commenting the inner loop and recompiling (thus enabling vectorization) reduces the runtime to 176.89s. Average timings for zhbmv in the four cases confirm the performance differences are due to this routine:

XT	XT libsci	XT netlib	XT netlib commented
2.4*10 <sup>-3</sup>	3.66*10 <sup>-2</sup>	1.97*10 <sup>-2</sup>	7*10 <sup>-4</sup>

The profiles also show that the single most time consuming routine is `mpi_all_to_all`. However, the performance of this routine is less significant on the X2 than on the XT as a percentage of the overall run time.

## CrayPAT Tracing

The sampling experiments listed above were used to construct tracing experiments using CrayPAT's Automatic Profiling Analysis option.

### XT allx1 16 proc

```
=====
USER / main
-----
Time%                67.3%
Time                 183.270900
Imb.Time             12.549153
Imb.Time%            7.3%
Calls                1
DATA_CACHE_MISSES   68.258M/sec  12368216931 misses
PAPI_TLB_DM          0.250M/sec   45337321 misses
PAPI_L1_DCA          1678.617M/sec 304161187441 refs
PAPI_FP_OPS          2162.653M/sec 391867404523 ops
User time (approx)   181.197 secs 471113500000 cycles
Average Time per Call 0.000000 sec/call
Overhead / Time      2267.3%
Cycles               181.197 secs 471113500000 cycles
User time (approx)   181.197 secs 471113500000 cycles
Utilization rate     100.0%
HW FP Ops / Cycles   0.83 ops/cycle
HW FP Ops / User time 2162.653M/sec 391867404523 ops 41.6%peak
HW FP Ops / WCT      2162.653M/sec
Computation intensity 1.29 ops/ref
MFLOPS               34602.46M/sec
LD & ST per TLB miss 6708.85 refs/miss
LD & ST per D1 miss  24.59 refs/miss
D1 cache hit ratio   95.9%
% TLB misses / cycle 0.0%
=====
```

### X2 Bespoke FFTs allx1 16 proc

```
=====
USER / castep_
-----
Time%                53.0%
Time                 179.718458
Imb.Time             5.246363
Imb.Time%            3.2%
Calls                1
VOPS_VL              3728.603M/sec 457400583884 ops
PAPI_VEC_INS         78.268M/sec   9601388532 instr
DCACHE_HIT           29.006M/sec   3558224331 hits
DCACHE_MISS          4.974M/sec    610157559 misses
PAPI_TOT_INS         330.429M/sec  40534828013 instr
PAPI_FP_OPS          2847.366M/sec 349296196196 ops
PAPI_TOT_CYC         122.673 secs  98138767428 cycles
User time (approx)   133.084 secs 106467465000 cycles
Average Time per Call 0.000000 sec/call
Overhead / Time      14929.7%
Cycles               122.673 secs  98138767428 cycles
User time (approx)   133.084 secs 106467465000 cycles
Utilization rate     100.0%
Instr per cycle       0.41 inst/cycle
HW FP Ops / Cycles 3.56 ops/cycle
=====
```

<b>HW FP Ops / User time</b>	<b>2847.366M/sec</b>	<b>349296196196 ops</b>	<b>11.1%peak</b>
HW FP Ops / WCT	2847.366M/sec		
HW FP Ops / Inst		861.7%	
<b>Avg VL</b>		<b>47.64 ops</b>	
Data cache refs	33.979M/sec	4168381890 refs	
D cache hit ratio		85.4%	
MIPS	5286.86M/sec		
MFLOPS	45557.85M/sec		
Instructions per LD ST		9.72 inst/ref	
LD & ST per D1 miss		6.83 refs/miss	

### X2 libsci FFTs allx1 16 proc

```
=====
USER / castep_
-----
---
```

Time%		63.6%	
Time		218.313417	
Imb.Time		5.040330	
Imb.Time%		2.6%	
Calls		1	
VOPS_VL	3017.798M/sec	498497777267 ops	
PAPI_VEC_INS	81.736M/sec	13501616871 instr	
DCACHE_HIT	43.989M/sec	7266421874 hits	
DCACHE_MISS	3.619M/sec	597870529 misses	
PAPI_TOT_INS	416.228M/sec	68755074792 instr	
PAPI_FP_OPS	2270.404M/sec	375038916911 ops	
PAPI_TOT_CYC	165.186 secs	132148763363 cycles	
User time (approx)	173.750 secs	138999970000 cycles	
Average Time per Call		0.000000 sec/call	
Overhead / Time		12868.7%	
Cycles	165.186 secs	132148763363 cycles	
User time (approx)	173.750 secs	138999970000 cycles	
Utilization rate		100.0%	
Instr per cycle		0.52 inst/cycle	
<b>HW FP Ops / Cycles</b>		<b>2.84 ops/cycle</b>	
<b>HW FP Ops / User time</b>	<b>2270.404M/sec</b>	<b>375038916911 ops</b>	<b>8.9%peak</b>
HW FP Ops / WCT	2270.404M/sec		
HW FP Ops / Inst		545.5%	
<b>Avg VL</b>		<b>36.92 ops</b>	
Data cache refs	47.609M/sec	7864292403 refs	
D cache hit ratio		92.4%	
MIPS	6659.65M/sec		
MFLOPS	36326.47M/sec		
Instructions per LD ST		8.74 inst/ref	
LD & ST per D1 miss		13.15 refs/miss	

```
=====
```

These tables give figures for the whole CASTEP executable, including calls to library routines that are hidden from CrayPAT, such as the libsci FFTs.

Comparing the XT table with X2 tables, an immediate difference is the peak performance figure – 41.6% on the XT and 11.1% and 8.9% on the X2. This suggests that CASTEP is using the hardware of the XT much better than that of the X2 in this case.

Comparing the bespoke and libsci FFT routines shows that the executable using bespoke FFTs makes better use of the vector registers:

- the average vector length is longer (47.6 vs. 36.2),
- the number of floating point operations per cycle is greater (3.56 vs 2.84). (The X2 is capable of retiring 16 floating point operations per cycle - e.g 8 multiply and 8 add - after the pipelines have been initialised.)
- the FLOPS rates are greater (running at 11.1% of theoretical peak vs. 8.9%)

Since the bespoke FFTs are user-supplied code we get a PAT report for these routines, and, for example, the following table shows output for the routine gpfa5f (actually rad5i and rad5ii, which are called from gpfa5f):

### X2 bespoke FFTs allx1 16 proc

```
=====
USER / gpfa5f$fft_gpfa_
-----
```

Time%		1.1%
Time		3.714851
Imb.Time		0.135285
Imb.Time%		4.0%
Calls		251680
VOPS_VL	4155.213M/sec	15495551976 ops
PAPI_VEC_INS	64.058M/sec	238883853 instr
DCACHE_HIT	36.236M/sec	135130645 hits
DCACHE_MISS	1.145M/sec	4271125 misses
PAPI_TOT_INS	231.495M/sec	863289163 instr
PAPI_FP_OPS	3062.210M/sec	11419543975 ops
PAPI_TOT_CYC	3.729 secs	2983346812 cycles
User time (approx)	0.403 secs	322450000 cycles
Average Time per Call		0.000000 sec/call
Overhead / Time		181781789943.4%
Cycles	3.729 secs	2983346812 cycles
User time (approx)	0.403 secs	322450000 cycles
Utilization rate		100.0%
Instr per cycle		0.29 inst/cycle
<b>HW FP Ops / Cycles</b>		<b>3.83 ops/cycle</b>
<b>HW FP Ops / User time</b>	<b>3062.210M/sec</b>	<b>11419543975 ops12.0%peak</b>
HW FP Ops / WCT	3062.210M/sec	
HW FP Ops / Inst		1322.8%
<b>Avg VL</b>		<b>64.87 ops</b>
Data cache refs	37.381M/sec	139401770 refs
D cache hit ratio		96.9%
MIPS	3703.93M/sec	
MFLOPS	48995.36M/sec	
Instructions per LD ST		6.19 inst/ref
LD & ST per D1 miss		32.64 refs/miss

```
=====
```

This table can be compared with the output for the rad5i routine on the XT:

### XT allx1 16 proc

```
=====
USER / RAD5I.in.FFT_GPFA
-----
```

Time%		2.4%
Time		6.715598
Imb.Time		0.096200

Imb.Time%		1.6%	
Calls		188725	
DATA_CACHE_MISSES	118.223M/sec	660403246	misses
PAPI_TLB_DM	0.048M/sec	269989	misses
PAPI_L1_DCA	1215.428M/sec	6789493566	refs
PAPI_FP_OPS	2227.874M/sec	12445115211	ops
User time (approx)	5.586 secs	14523843750	cycles
Average Time per Call		0.000000	sec/call
Overhead / Time		11505329430.4%	
Cycles	5.586 secs	14523843750	cycles
User time (approx)	5.586 secs	14523843750	cycles
Utilization rate		100.0%	
HW FP Ops / Cycles		0.86	ops/cycle
<b>HW FP Ops / User time</b>	<b>2227.874M/sec</b>	<b>12445115211</b>	<b>ops 42.8%peak</b>
HW FP Ops / WCT	2227.874M/sec		
Computation intensity		1.83	ops/ref
MFLOPS	35645.99M/sec		
LD & ST per TLB miss		25147.33	refs/miss
LD & ST per D1 miss		10.28	refs/miss
D1 cache hit ratio		90.3%	
% TLB misses / cycle		0.0%	

=====  
Although the vector machine is processing more floating-point instructions per second (3.06 GFLOPS vs. 2.2 GFLOPS), this is somewhat short of the theoretical peak performance of the X2 (25.6 GFLOPS); the XT run achieves much closer to the theoretical peak of a single Opteron core (5.6 GFLOPS).

The results for the al1x1 16 proc case may be compared with those for the larger al3x3 16 proc case.

### X2 bespoke FFTs al3x3 16 proc

=====  
USER / castep\_  
-----  
---

Time%		84.3%	
Time		3942.380814	
Imb.Time		52.875652	
Imb.Time%		1.5%	
Calls		1	
VOPS_VL	10786.848M/sec	41787189905570	ops
PAPI_VEC_INS	118.336M/sec	458423644048	instr
DCACHE_HIT	24.275M/sec	94038801549	hits
DCACHE_MISS	1.897M/sec	7349945411	misses
PAPI_TOT_INS	312.953M/sec	1212349784141	instr
PAPI_FP_OPS	9517.049M/sec	36868112621489	ops
PAPI_TOT_CYC	3873.902 secs	3099121303275	cycles
User time (approx)	3899.569 secs	3119655485000	cycles
Average Time per Call		0.000005	sec/call
Overhead / Time		669.8%	
Cycles	3873.902 secs	3099121303275	cycles
User time (approx)	3899.569 secs	3119655485000	cycles
Utilization rate		100.0%	
Instr per cycle		0.39	inst/cycle
<b>HW FP Ops / Cycles</b>		<b>11.90</b>	<b>ops/cycle</b>
<b>HW FP Ops / User time</b>	<b>9517.049M/sec</b>	<b>36868112621489</b>	<b>ops37.2%peak</b>
HW FP Ops / WCT	9517.049M/sec		
HW FP Ops / Inst		3041.0%	
<b>Avg VL</b>		<b>91.15</b>	<b>ops</b>



Data cache refs	26.172M/sec	101388746959 refs
D cache hit ratio		92.8%
MIPS	5007.25M/sec	
MFLOPS	152272.79M/sec	
Instructions per LD ST		11.96 inst/ref
LD & ST per D1 miss		13.79 refs/miss

=====  
This shows a considerable improvement over the smaller allx1 case. This larger case will involve greater loop bounds (demonstrated by an average vector length increase from 64.87 to 91.15) and therefore make more sustained use of the vector unit, resulting in a good 11.9 FP ops per cycle and running at an acceptable 37.2% of peak.

### X2 bespoke FFTs al3x3 16 proc

=====  
USER / gpfa5f\$fft\_gpfa\_  
-----  
---

Time%		1.1%
Time		49.389664
Imb.Time		1.689171
Imb.Time%		3.8%
Calls		1239018
VOPS_VL	5690.326M/sec	272427152070 ops
PAPI_VEC_INS	46.389M/sec	2220891576 instr
DCACHE_HIT	12.135M/sec	580972618 hits
DCACHE_MISS	0.537M/sec	25687919 misses
PAPI_TOT_INS	122.806M/sec	5879394154 instr
PAPI_FP_OPS	4193.054M/sec	200744517417 ops
PAPI_TOT_CYC	47.875 secs	38300393033 cycles
User time (approx)	44.244 secs	35395230000 cycles
Average Time per Call		0.000000 sec/call
Overhead / Time		66245897039.1%
Cycles	47.875 secs	38300393033 cycles
User time (approx)	44.244 secs	35395230000 cycles
Utilization rate		100.0%
Instr per cycle		0.15 inst/cycle
<b>HW FP Ops / Cycles</b>		<b>5.24 ops/cycle</b>
<b>HW FP Ops / User time</b>	<b>4193.054M/sec</b>	<b>200744517417 ops 16.4%peak</b>
HW FP Ops / WCT	4193.054M/sec	
HW FP Ops / Inst		3414.4%
<b>Avg VL</b>		<b>122.67 ops</b>
Data cache refs	12.672M/sec	606660536 refs
D cache hit ratio		95.8%
MIPS	1964.89M/sec	
MFLOPS	67088.86M/sec	
Instructions per LD ST		9.69 inst/ref
LD & ST per D1 miss		23.62 refs/miss

=====

The improved performance in the FFT routine can again be attributed to the fact that the loop bounds are larger (122.67 average vector length) and therefore the vector unit can be kept busier for longer. However, the loops in rad5i are such that they will not keep both the add and multiply units busy all the time, and so the peak ops per cycle rate will never get close to 16; a figure of 5.24 is acceptable.

## **Conclusion**

The limiting factor for the performance of CASTEP on the X2 seems to be the performance of the BLAS, which causes a particular problem in the NaHF2 case. The performance of the bespoke CASTEP FFTs is consistently better than that of the libsci FFTs. The a13x3, 16 proc. case appears to be making best use of the vector machine. This is probably due to partitioning more work per processor and therefore sustaining more activity in the vector unit, thus minimising vector pipeline start-up costs. This indicates that CASTEP makes acceptable use of the X2 machine for cases where each processor is allocated enough work to keep the vector unit busy.