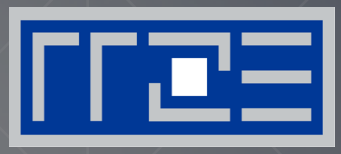
ERLANGEN REGIONAL COMPUTING CENTER

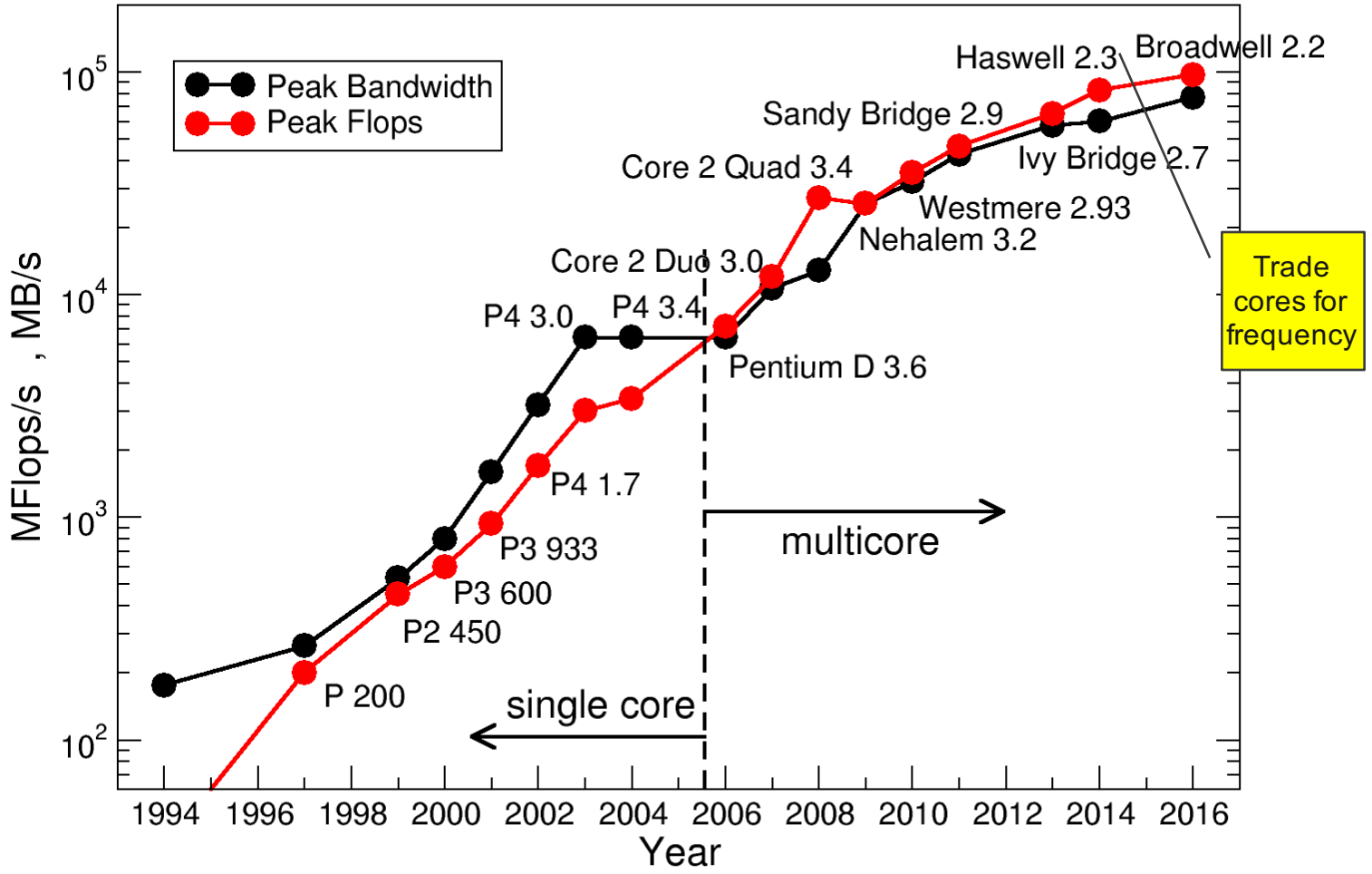


Evaluation of Intel Xeon Phi "Knights Landing": Initial impressions and benchmarking results

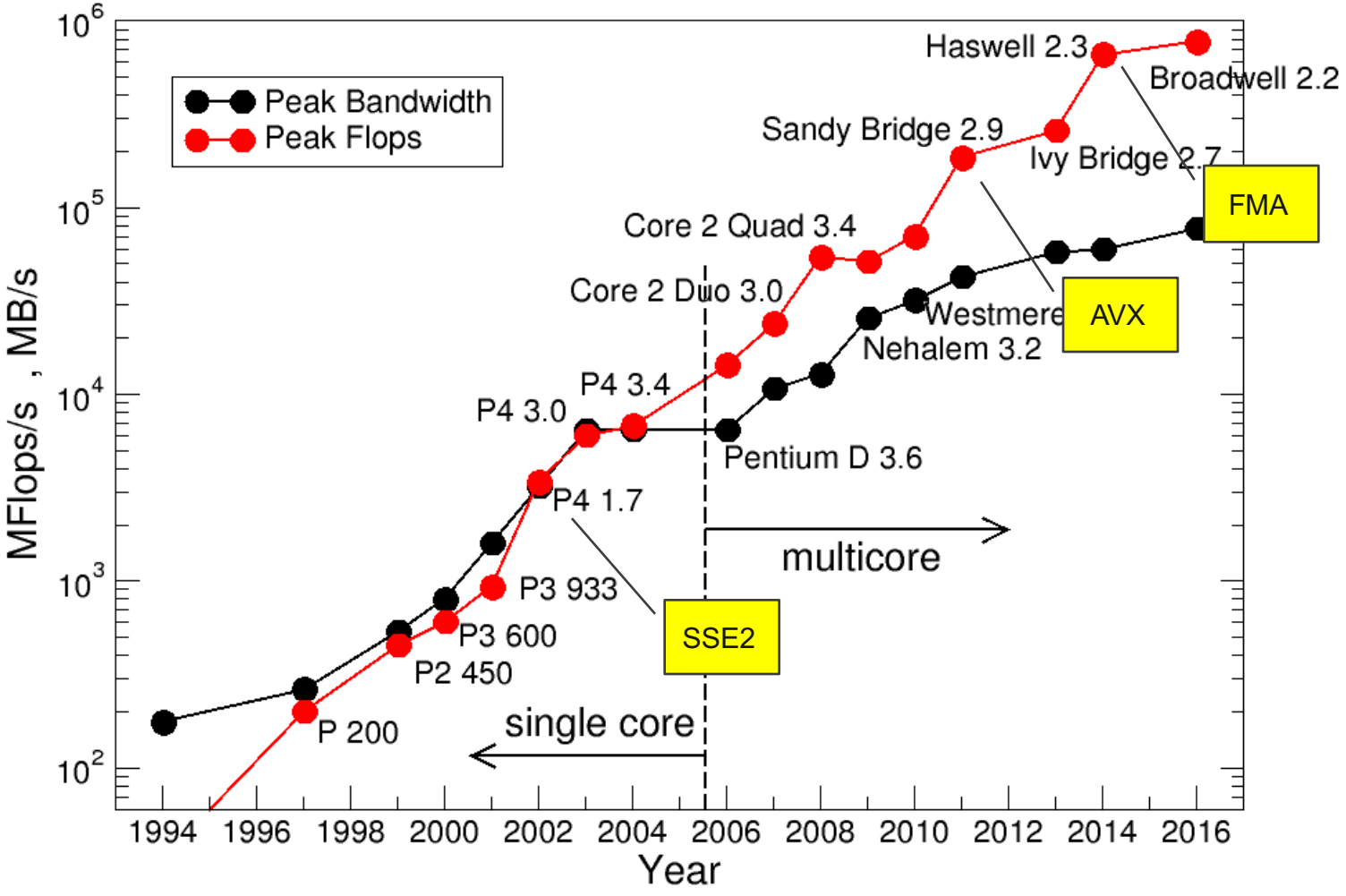
J. Eitzinger

PRACE PATC, 28.6.2017

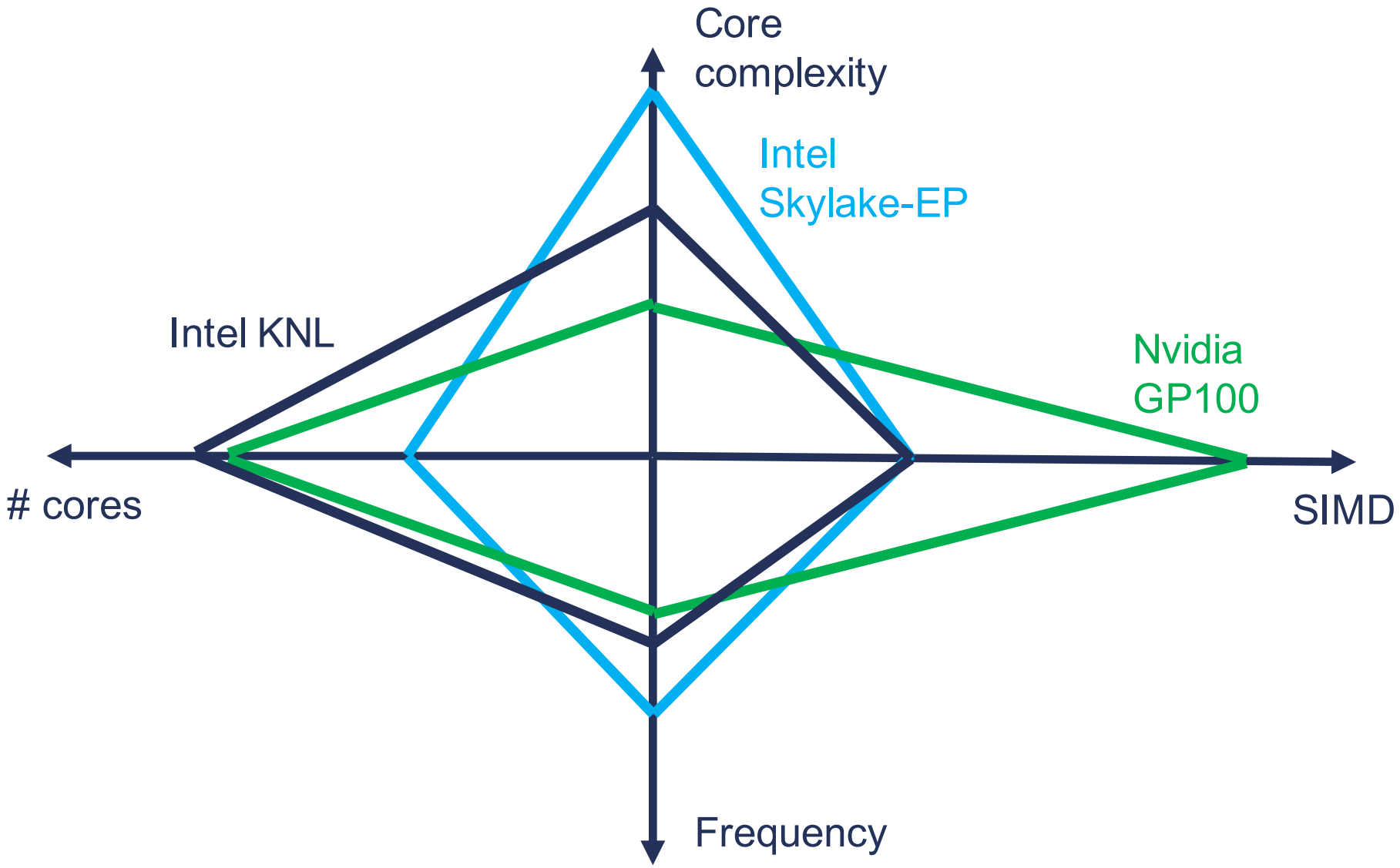
History of Intel hardware developments



The real picture



Finding the right compromise



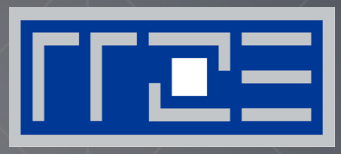
Maximum DP floating point (FP) performance

$$P_{core} = n_{super}^{FP} \cdot n_{FMA} \cdot n_{SIMD} \cdot f$$

- n_{super}^{FP}: Super-scalarity
- n_{FMA}: FMA factor
- n_{SIMD}: SIMD factor
- f: Clock Speed

uArch	n_{super}^{FP}	n_{FMA}	n_{SIMD}	n_{cores}	Release	Model	P_{core} [GF/s]	P_{chip} [GF/s]	P_{serial} [GF/s]	TDP	GF/ Watt
Sandy Bridge	2	1	4	8	Q1/2012	E5-2680	11.7	173	7	130	1,33
Ivy Bridge	2	1	4	10	Q3/2013	E5-2690-v2	24	240	7,2	130	1,85
KNC	1	2	8	61	Q2/2014	7120A	10.6	1210	1,33	300	4,03
Haswell	2	2	4	14	Q3/2014	E5-2695-v3	21.6	425	6,6	120	3,54
Broadwell	2	2	4	22	Q1/2016	E5-2699-v4	17.6	704	7,2	145	4,85
Pascal	1	2	32	56	Q2/2016	GP100	36.8	4700	1,5	300	15,67
KNL	2	2	8	72	Q4/2016	7290F	35.2	2995	3,4	260	11,52
Skylake	2	2	8	26	Q3/2017	8170	23.4	1581	7,6	165	9,58

ERLANGEN REGIONAL COMPUTING CENTER



Chebyshev Filter Diagonalization on KNL and P100

DFG SPPEXA Essex 2 project

Basic ChebFD scheme

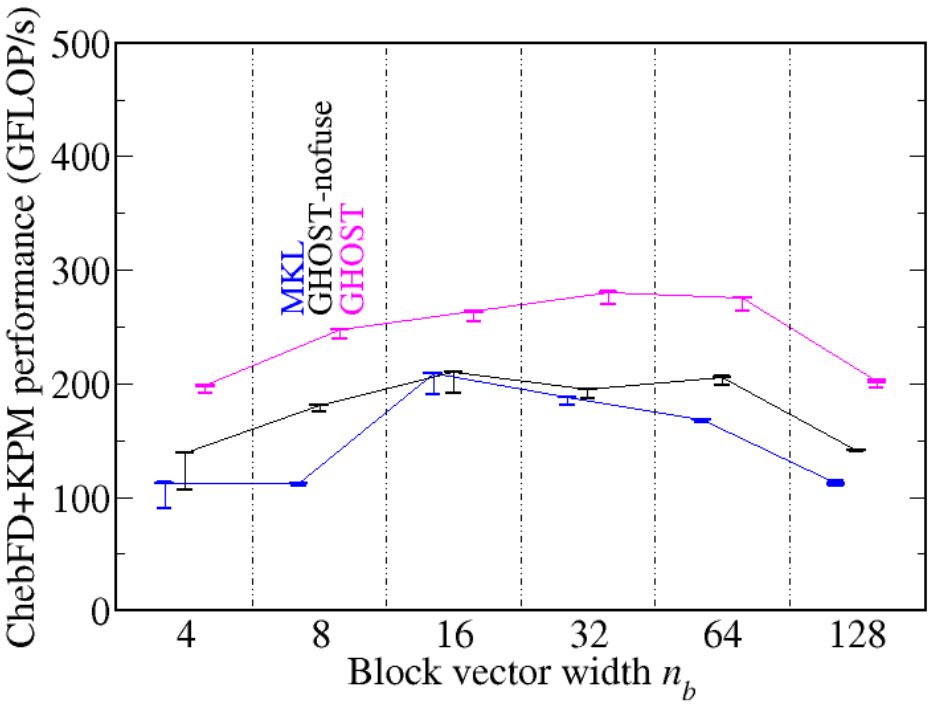
1. Filter n_s search vectors
2. Orthogonalize n_s search vectors
3. Go to 1 if not converged

- n: matrix/vector dimension
- n_p: polynomial degree (defined by application)
- n_s: number of search vectors (defined by application)

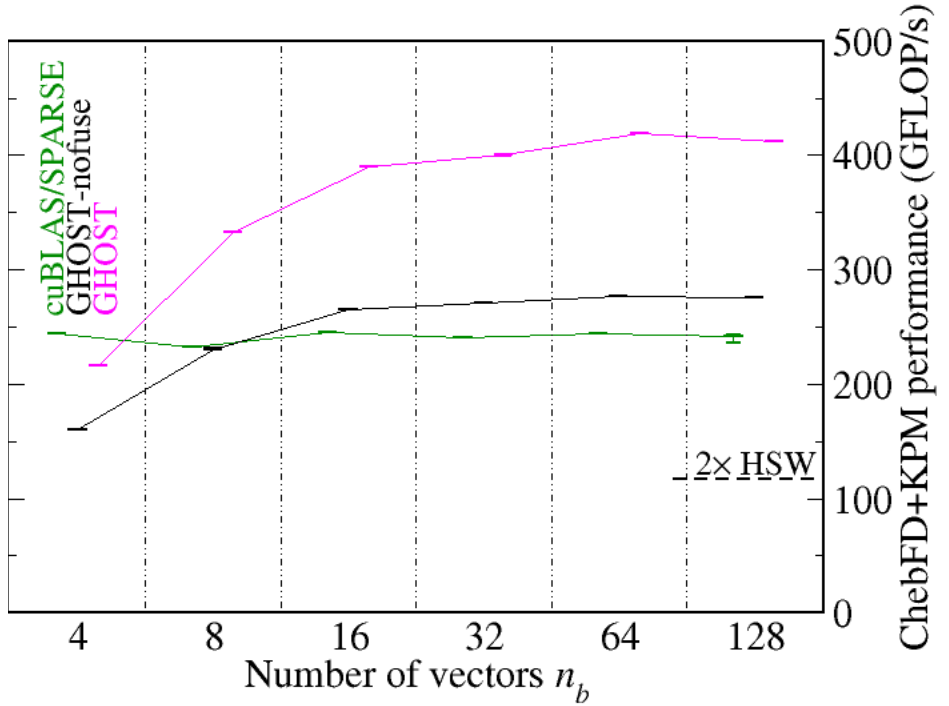
Test systems

- **Piz Daint** (Switzerland): 5320 Node Cray XC50
- **Oakforst-PACS** (Japan): 8208 Node Fujitsu PRIMERGY

Node-level Performance

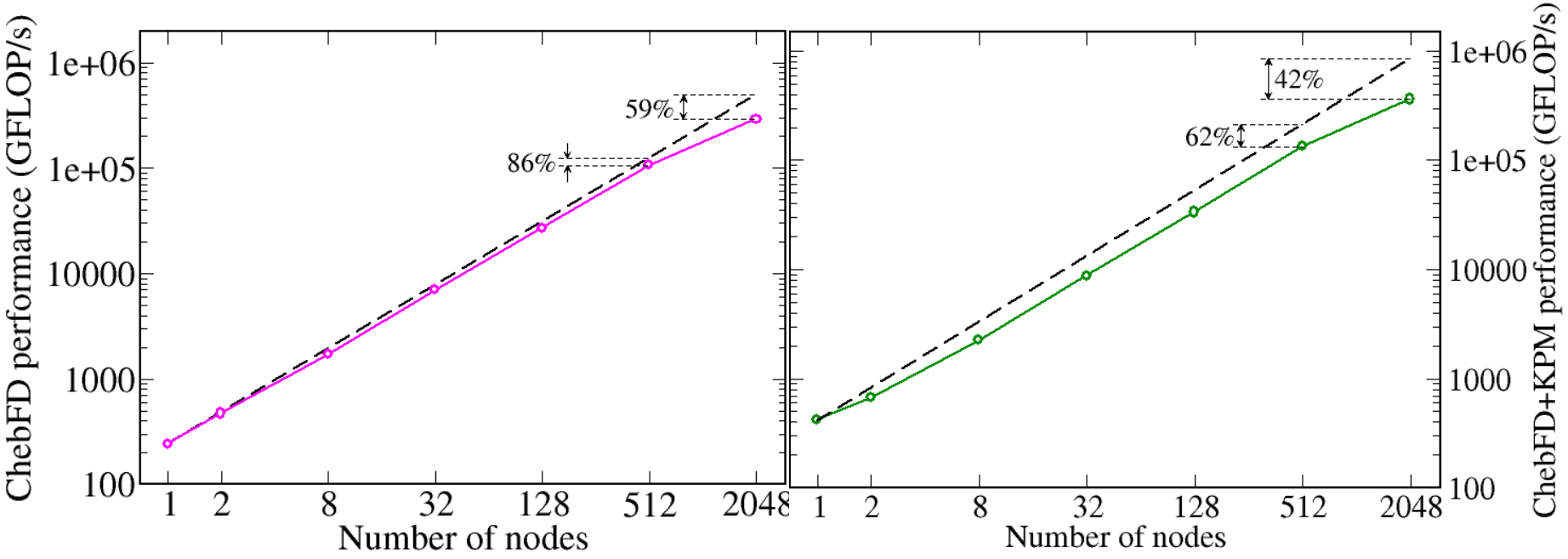


KNL



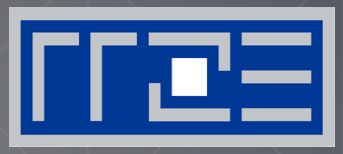
P100

Scaling Results



Fujitsu PRIMERGY

Cray XC50



System configuration challenge

Configuration complexity

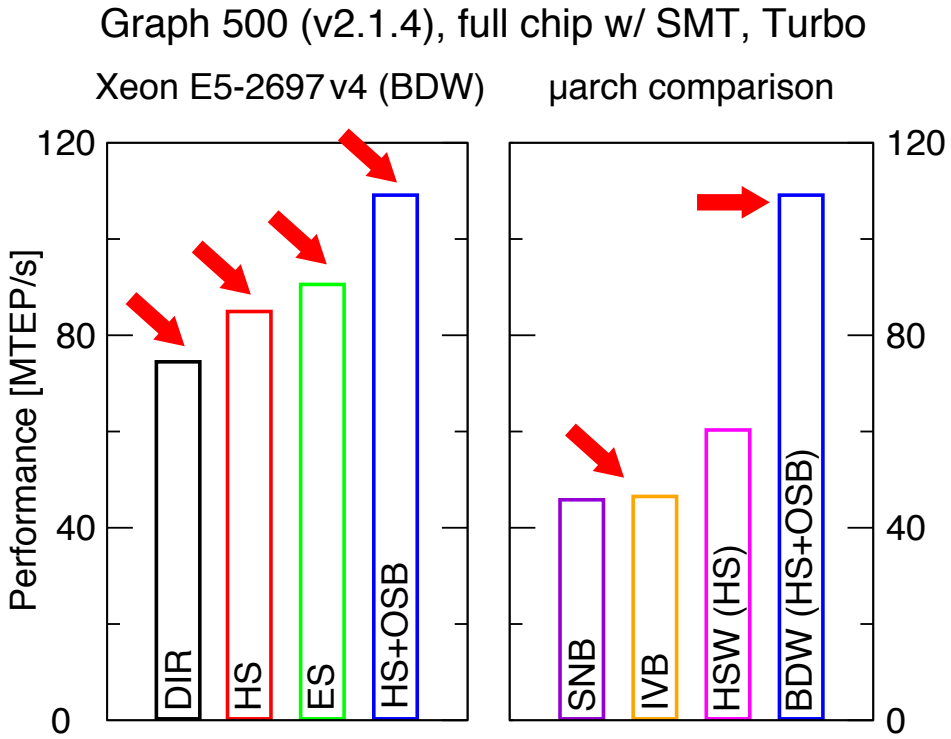
- **Cluster modes**: lower the latency and increase the bandwidth
  - All-to-all
  - Quadrant mode (default)
  - Sub-numa-clustering (SNC), best performance but explicit
- **Memory modes**:
  - Cache mode (default)
  - Flat mode (explicit)
  - Hybrid
- **Mapping** of application on hardware:
  - Use SMT or not. How many SMT threads?
  - Use all cores?
  - MPI+X. How exactly?
- **Memory configuration**: Alignment and page size choices

Impact of QPI snoop mode and CoD on latency

- Starting with HSW, QPI snoop mode can be set via BIOS
  - Early Snoop
  - Home Snoop
  - Home Snoop + Opportunistic Snoop Broadcast (BDW only)
  - Directory (CoD)

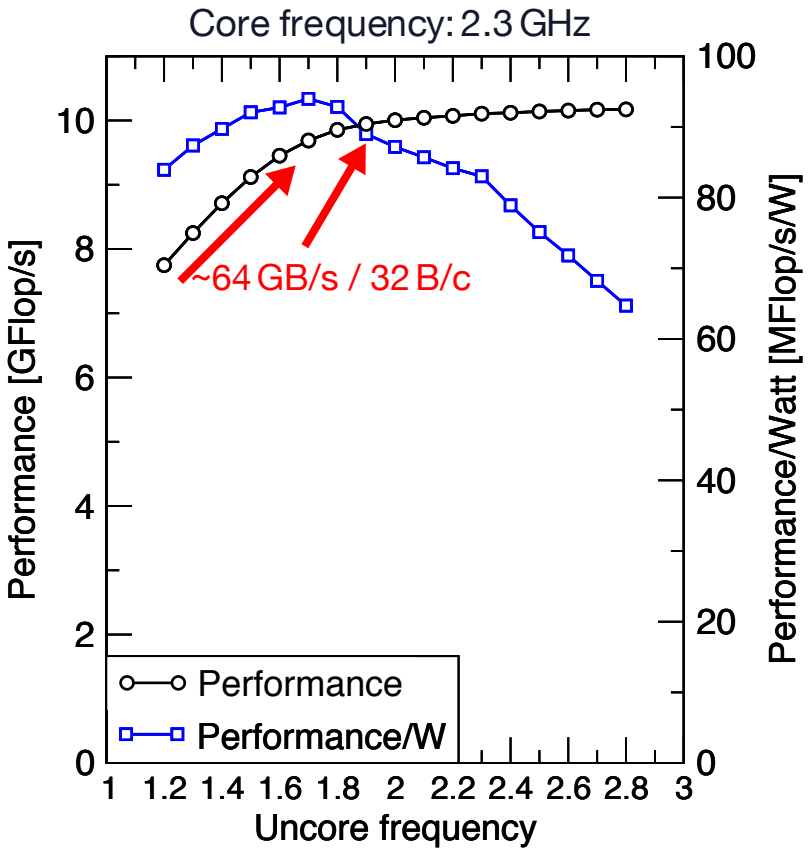
	SNB	IVB	HSW	BDW
L1	4	4	4	4
L2	12	12	12	12
L3	40	40	37 (COD)	41 (CoD) 47 (non-CoD)
Mem	230	208	168 (COD)	280 (HS), 248 (ES), 190 (HS+OSB), 176 (DIR)

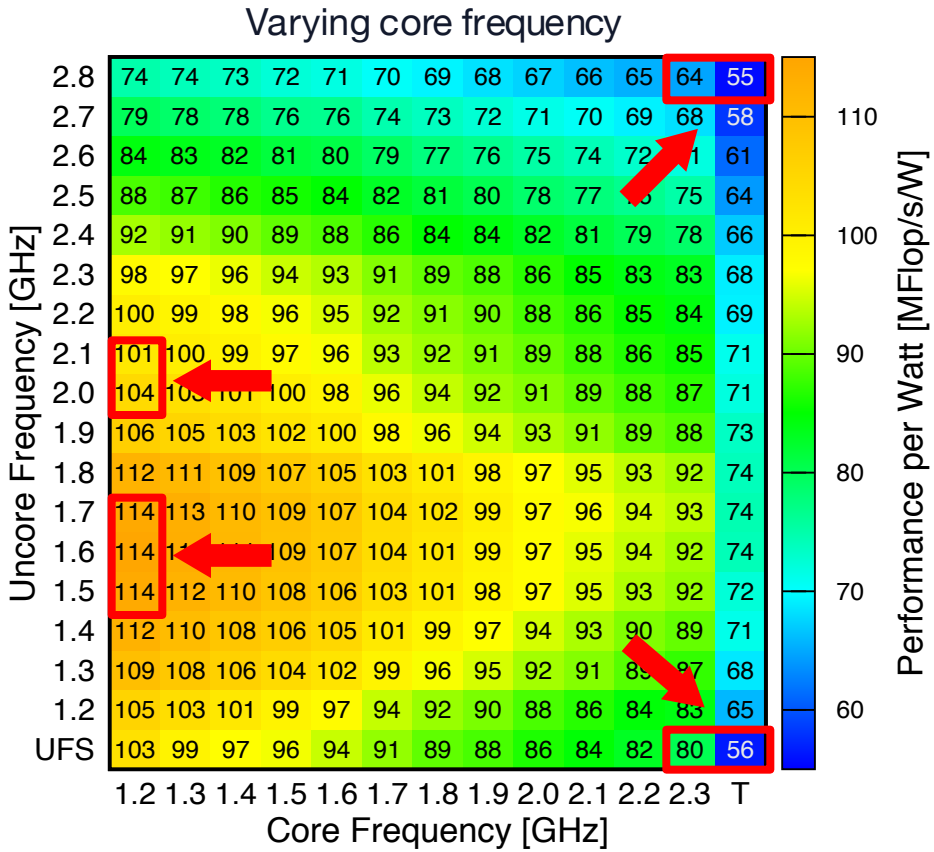
Cache/Memory Latency [cycles]



Uncore Frequency: Bandwidth and Energy Consumption

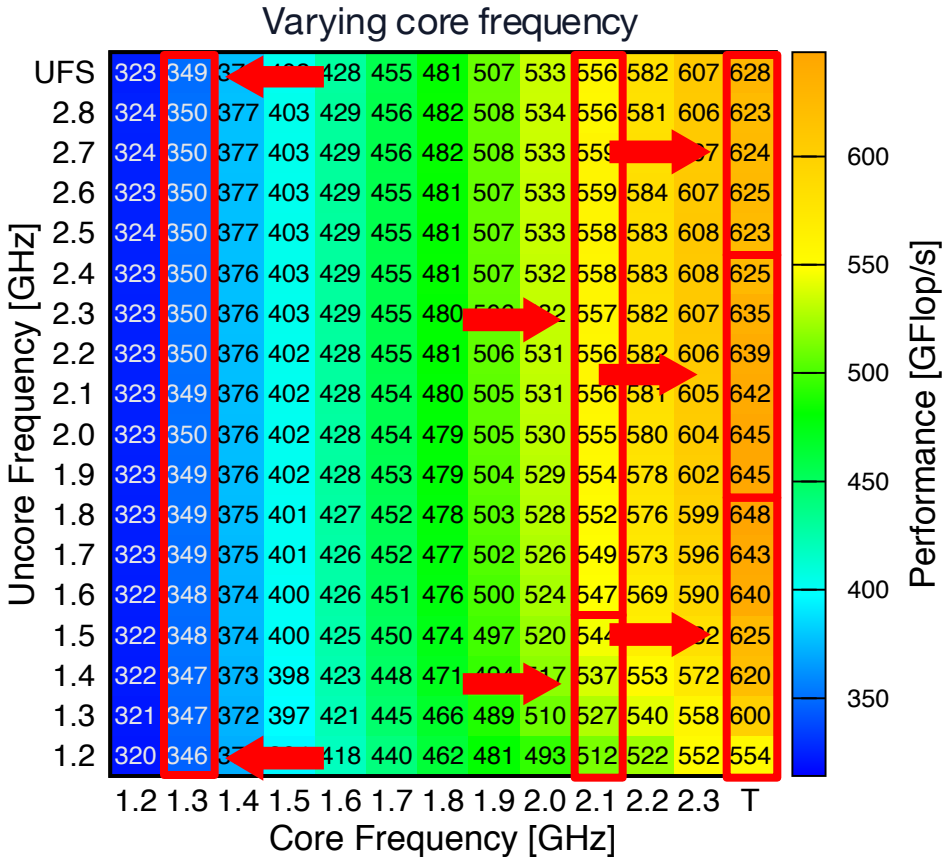
Intel HPCG (16.0.3), n=256, full chip (no SMT), Xeon E5-2697 v4 (BDW)

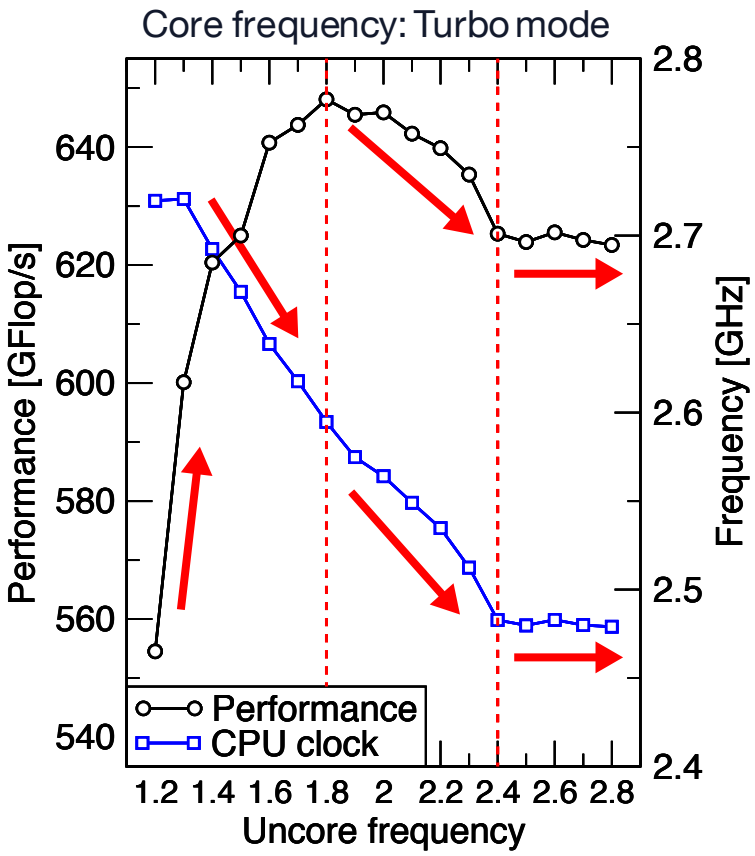




Uncore Frequency: LINPACK Performance

Intel HPL (16.0.3), N=60.000, full chip (no SMT), Xeon E5-2697 v4 (BDW)





Specific issues with Xeon Phi

- **MCDRAM** adds additional complexity
- **Configuration** of system and **mapping** of application on hardware gets more critical
- The compromise and made with KNL will soon be outdated
- KNL as a hosted cluster system is probably too specialized for a general purpose academic cluster

But

- Xeon Phi implements features which are not available anywhere else:
  - High degree of chip level parallelism
  - Multiple memory types and explicit memory control
  - Mesh type on-die topology

- It allowed a glimpse in the future on real hardware