



# EE382N-20: Intel Xeon Phi

Mattan Erez

The University of Texas at Austin



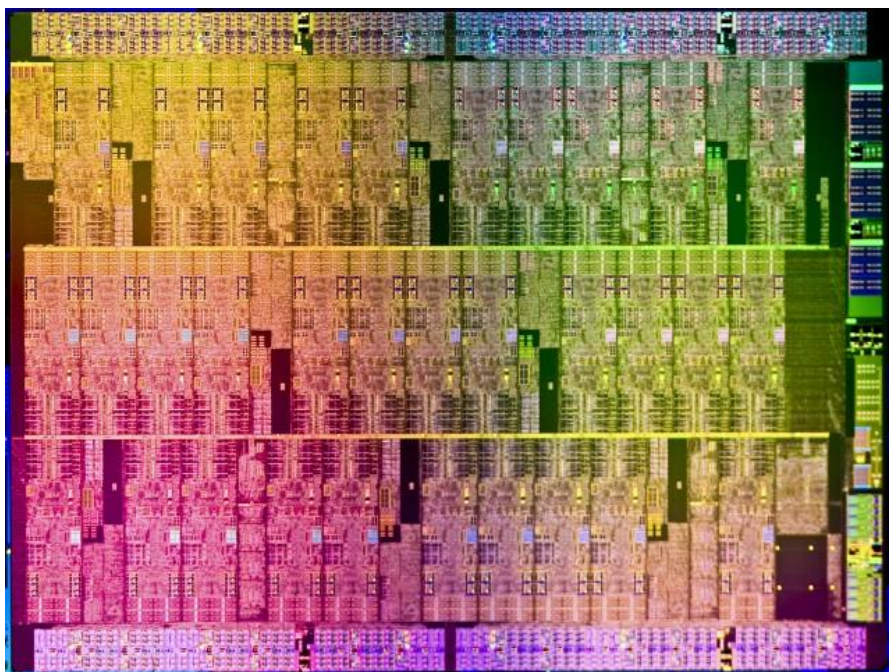
# Intel's version of throughput computing

- Similar goals as GPUs
- Ignore latency
- Focus on parallelism and efficiency
- But, start with traditional x86 cores
  - More general purpose
  - Reasonable caches
- Add more parallelism and more latency hiding



# Knights Corner

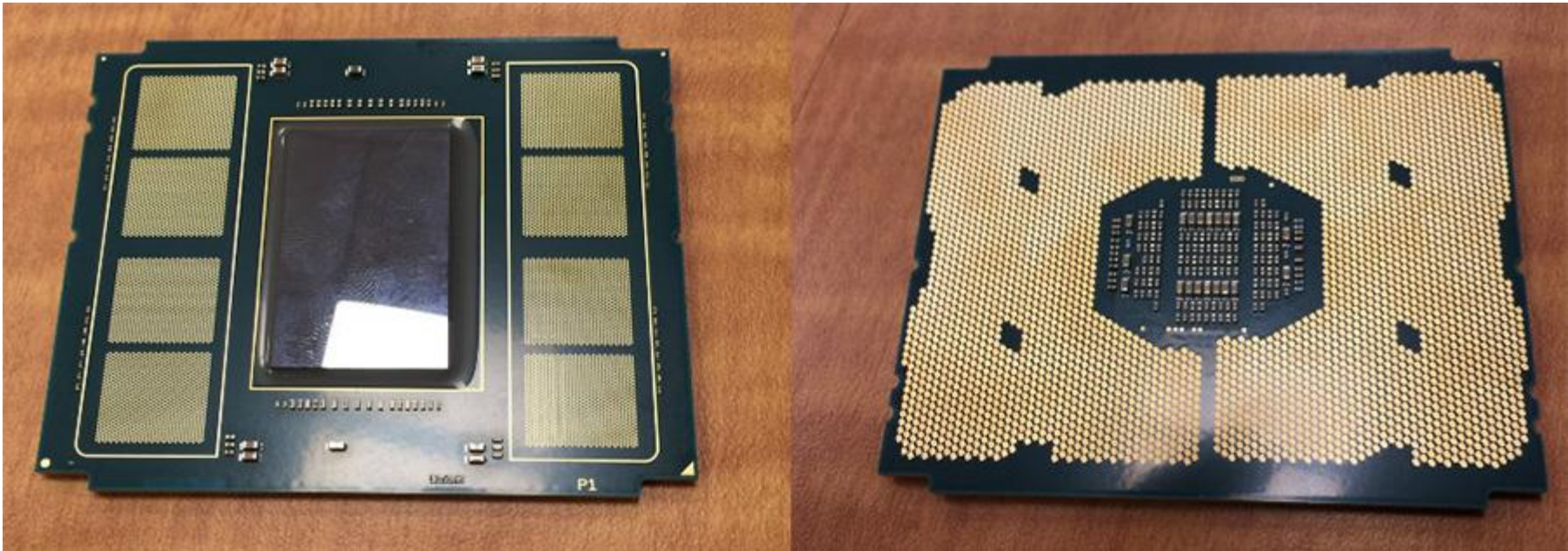
- Pentium (P54) scalar cores
  - Efficient small core, but not best performing
- 61 cores (22nm process)
- 512KB shared L2 per core





# Knights Landing

- Upgrade to Atom cores
- More cores (72) (14nm process)
- On-package “near” DRAM
- Socket rather than PCIe
  - Can be “self hosted”





# Parallelism

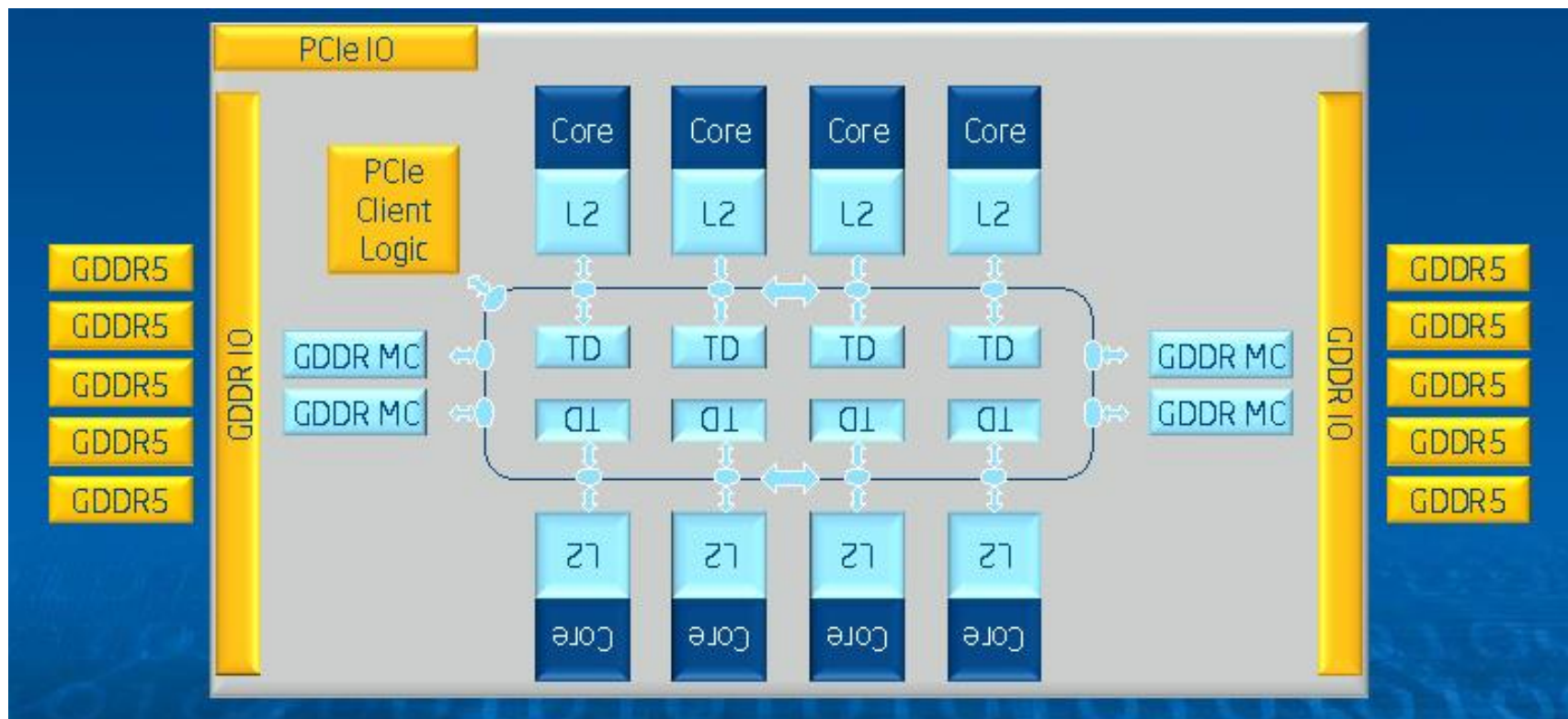
- Forget OOO, explicit parallelism
- 512-bit vector units in each core
- “many integrated cores” (MIC)
  - Currently, 61 cores and more soon
  
- A lot of the innovations are in the SIMD units
  - New instructions
  - Better arithmetic design
  - Scatter/gather (to cache)
  - Help with mask generation
  - Better permutations





# Latency hiding

- 4 HW threads per core
- Many cores
- Big-ish caches





# Simplicity first

- Ring interconnect
- Extended MESI (kind of like MOESI, but not quite)
  - MESI + GOLS (globally owned locally shared)
- Simple global Tag Directory
- No real multi-threading support, just coherence
  - Currently, synchronization is painful
  - ~250 cycles to get through tag directory to find stuff



# Programming

- Currently, accelerator as off-load mode
- Program with OpenMP/OpenACC or MPI for the most part
- Can share virtual addresses with host and compiler can insert copies (at offload boundaries)
- Still a bit rough around the edges
  - Performance bugs
  - Memory leaks
  - Inter-node comm not ideal
- Significant investment though, will get better



