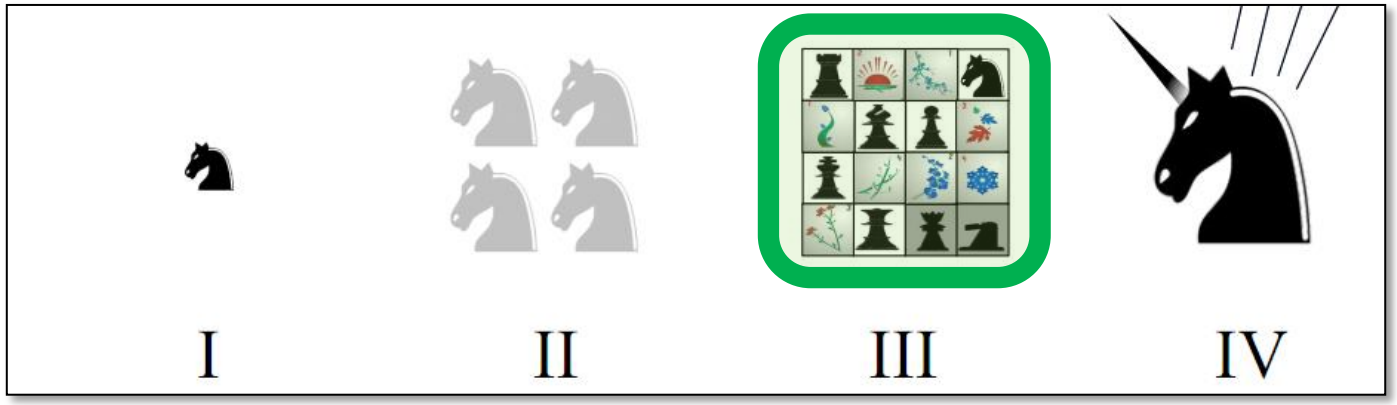
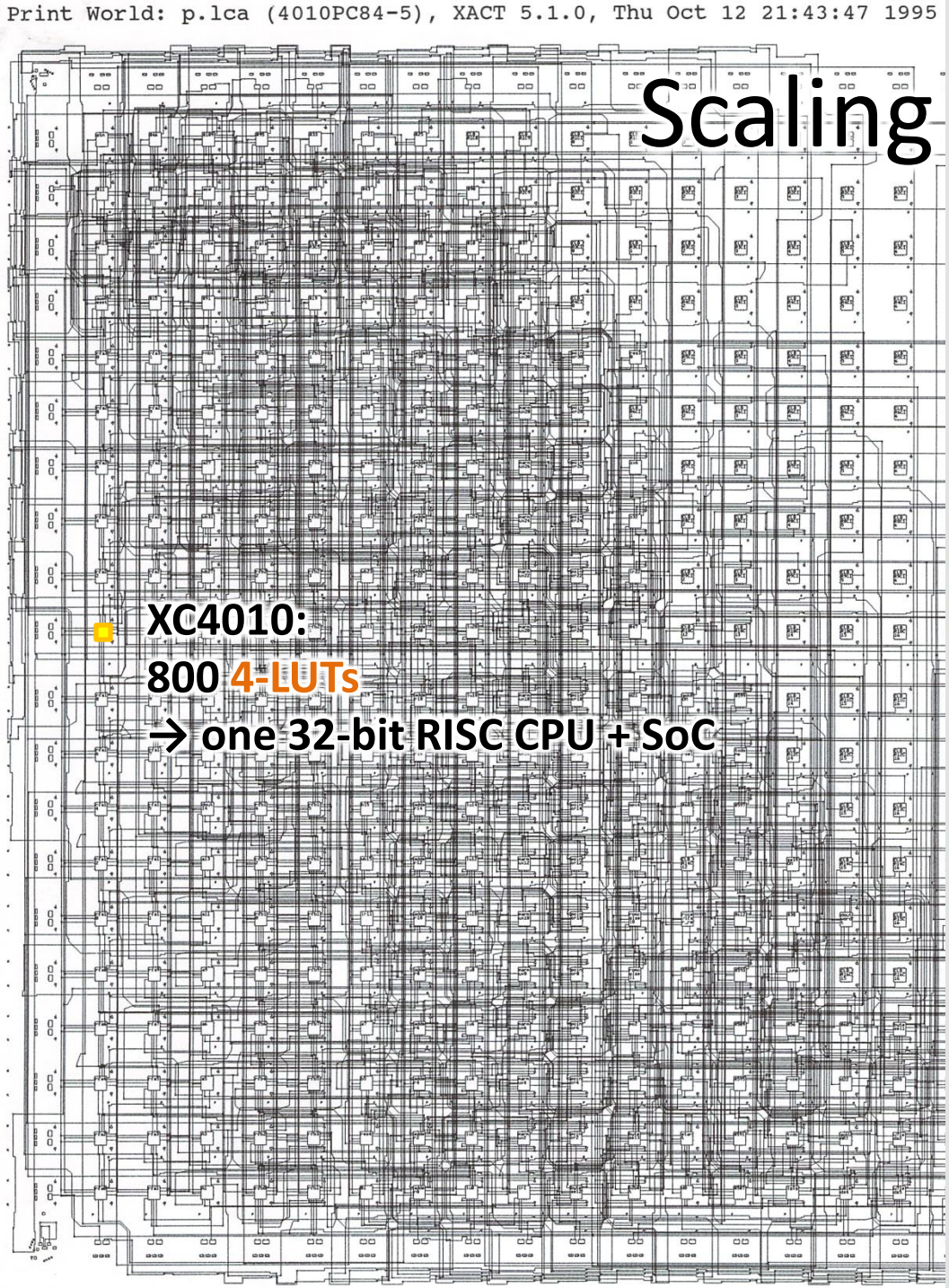
Dark Silicon?
Hot, Modular, Specialized, Integrated, Heterogeneous Silicon

Jan Gray
Gray Research LLC

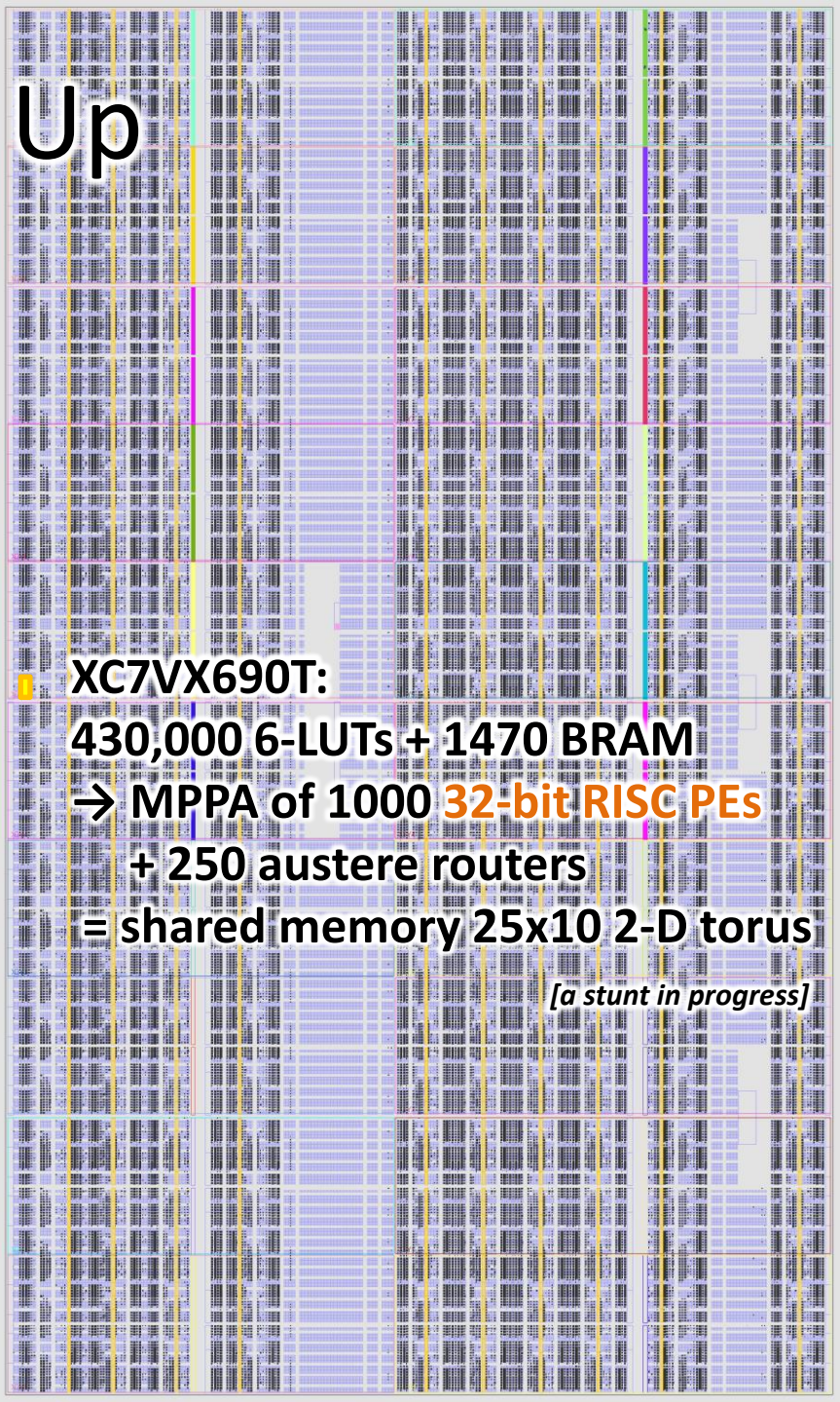


"The Four Horsemen of the Coming Dark Silicon Apocalypse"
Michael Taylor, UCSD. [http://darksilicon.org/horsemen/horsemen_slides.pdf]

Scaling Up



- XC4010:
  800 4-LUTs
  → one 32-bit RISC CPU + SoC



- XC7VX690T:
  430,000 6-LUTs + 1470 BRAM
  → MPPA of 1000 32-bit RISC PEs
  + 250 austere routers
  = shared memory 25x10 2-D torus

[a stunt in progress]

FPGAs: What's Next?

- 2013
  - 28 nm low-power
  - 600 KLUTs
  - 2.5D SSI
    - 1.2 MLUTs
    - 28 Gbps serdes (hetero)
  - + 2xARM A9s+SoC
    - ARM software tools, AXI IP cores portfolio
  - Vivado, HLS, OpenCL
  - Ubuntu, Android
  - New consumer apps: 4K TV, automotive, CV
- 2017
  - 14 nm *lower-power*, but  ?

    http://en.wikipedia.org/wiki/File:Robert_Dennard.jpg
  - 2+ MLUTs
  - 2.5D SSI
    - 4+ MLUTs
    - HMC DRAM, 50++ Gbps serdes
  - ARM-SoCs in almost every "FPGA"
    - *"Virtex-9* ***Datacenter Edition****"*? HOT CHIPS
      8+ A57 + SoC + FPGA + HMC + serdes + …
  - + … + DSLs + OpenCL/HLS libraries
  - Android, Windows, AArch64 CentOS, Windows Server Datacenter
  - New datacenter apps? SDN, ML, CV
  - *Fully used → same thermals issues. Dark Fabric?*

Computing Goes Specialized, Integrated, Heterogeneous

- Continued performance scaling = ↓↓↓energy/computation
- Cut waste:
  - Complex → essential
  - CPUs → specialized accelerators, memories, interconnects
  - Rethink, collapse legacy abstraction layers *SW*↔*SW*↔*HW*↔*HW*
  - Provision right
- Example: datacenter workloads → 2.5D-FPGA-SOCs?
  - Xeons + DRAM→ ARMs + <u>parallel FCCMs + HMC</u> (+ DRAM)
  - PCIe + 10G NICs + top of rack switches → FPGA-serdes i.c.
- Hot is OK

Challenges/Deal Breakers

- Mostly same scaling rules apply
- *Configurability cost* vs. ASICs, GPUs, mGPUs, Phi
- Poor tools – light years from software experience
- Missing interfaces and infrastructure

A Way Forward?

- Embrace SW models, languages, libs
- Build new infrastructure:
  - Common NOC and host interfaces
  - IP: router; CPUs; accelerators; NIC; DRAM (all speak NOC)
  - "OpenCL" → ARM, soft CPUs, accel'rs
  - "HeMPPAA" composer →
- App optimization: CPU → CPUs → soft CPUs → specialized CPUs→ accelerators
- More 10s recompiles; fewer 1h PARs; zero months interfacing to memory, host, OS; portable reusable code; composition
- ***Feasible! But…***