The raw prototype design document

by Yunqi Zhang
04/09/2013

1. Motivation
    a. wire delay: short wire, no longer than the size of one tile
    b. scalability: physical scalable, constant size with growing larger transistor budgets
    c. low design and verification complexity: replicate a single tile in scale
2. Design decisions:
    a. FPGA is simple, physically scalable parallel fabric
    b. FPGA allow fast communication and synchronization between entities
    c. FPGA is effective at bit and byte-wide data manipulation
    d. processors are highly optimized for datapath oriented computations
    e. compilation times are in seconds, not hours
3. Architecture:
    a. A raw tile: tile processor, static switch, dynamic router
    b. tile processor: 32-bit MIPS instruction set
    c. static switch processor: MIPS-like instruction set, contains just moves, branches and jumps
    d. dynamic router: independently running, under user control only indirectly
4. Static network design
    a. flow-control policy
        i. when the sequencer attempts to route a dataword which has not arrived yet, it will stall until it does arrive
        ii. provides support for programs with unpredictable timing
    b. static input block requires a small amount of buffering to prevent delay
5. Dynamic network primer
    a. limitations of static network
        i. the destinations must be known at compile time
        ii. the message size must be known at compile time
        iii. compiler needs to generate a switch schedule which merges 2 patterns on a cycle by cycle basis for cross communication routes
6. Tile processor design
    a. directly attached static network interface to the processor pipeline in order to minimize the latency from tile to tile
    b. input: map the network port names into the register file name space
    c. output: set the “S” bit to send the value out of the network and keep it locally
7. Deadlock
    a. high-priority network: “Matt Frank protocol”
        i. all users must obey it to ensure deadlock-free behaviour

ii. used for memory, interrupt, I/O, DMA and other communications go off-chip
    1. scale down the memory system
    2. partner memories: each tile is assigned to a particular DRAM
    3. memory dropbox: a DRAM all the tiles can access directly
    4. memory maintainer: at least one tile can access all of the memories

b. low-priority network: deadlock recovery