University of California at Berkeley
Department of Electrical Engineering and Computer Science
Reconfigurable Computing Seminar

CS294-7, Spring 1997 | Terminology | Tuesday, January 21

Terminology

This handout summarizes some of the terminology which we expect to use during the course and which may be unfamiliar to you. Please, use this as a first line of defense against confusion. If you find us using other terms which are unfamiliar to you or used in an unfamiliar manner, please let us know so we can clear up any confusion and expand this list.

N.B. The first draft of this list appeared as an appendix in Andre's thesis. [DeH96]. Unless otherwise noted, the term is used in his thesis and the terminology appendix printed there is likely to have references to the most relevant portions.

τ see tau.

λ see lambda.

ACS Adaptable Computing Systems – Term used by DARPA for the broad reconfigurable computing field. See `<http://www.ito.darpa.mil/ResearchAreas/Adaptive_Computing_Systems/ACSintro.html>`.

active computing resources The portions of a general-purpose architecture which actually compute results or transport data – *e.g.* ALUs, switches, wires. The term is typically used to distinguish such resources from overhead resources used to store descriptions or intermediate data.

active interconnect Switches and wires which actually produce a physical connection between a source and a destination. The term is used to distinguish resources used to actually perform switching from descriptions of switching operations or storage for intermediate data.

bit processing element A generic term for the primitive computational unit which produces one bit of result. Conventionally, each FPGA LUT is a bit processing element, as is each bit-slice in an SIMD ALU datapath.

CCM Custom Computing Machine – A generic term used to refer to configurable computing machines, usually built out of FPGAs and memories.

context A generic term used to refer to a slice of instructions and intermediate data used by a general-purpose device on a single cycle. See configuration context and data context.

control stream An independent thread of execution. When the computation varies with time and data, the control stream determines which sets of instructions are executed on a give cycle. A computational device may support a single control stream (*e.g.* processors, SIMD, pure VLIW) or multiple control streams (*e.g.* MSIMD, MIMD).

configurable computing Computing by configuring interconnect between programmable function units to wire up computations spatially.

configurable computing architectures Architectures where there is only one or a few instructions loaded per active computing element and there is limited bandwidth to reload an entire configuration context. These architectures are used for configurable computing where the computation is typically arranged via spatial interconnect of computing elements as opposed to programmable computing architectures which realize computation by rapid temporal reuse of a few, central active computing resources.

computational density See functional density.

computational throughput Computations performed per unit time. *i.e.* Operations completed per unit time.

configuration context The collection of bits which describe the behavior of a general-purpose machine on one operation cycle. Equivalently, the collection of all instructions required to specify the behavior of a general-purpose device at one point in time.

data context The data used by a general-purpose device on one cycle of execution.

distance delay The critical path delay through a placed circuit taking into account the distance between logically adjacent functional units.

datapath granularity Datapath width. The number of bit processing elements or interconnect switches controlled in SIMD fashion by a single instruction.

deployable resources Resources whose role can be determined at run-time. *e.g.* A memory which can be used as an instruction store or as a data store; Interconnect which can be used to distribute instructions or to deliver data between functional units. Distinguished from resources which are dedicated to a single function at fabrication time.

dynamic Marked by a continuous usually productive activity or change. In this context usually used to distinguish quantities, particularly, instructions, which change on a cycle-by-cycle basis. Contrast with static and quasistatic.

dynamic instruction distribution Instruction distribution allowing instructions to change on a cycle-by-cycle basis.

DPGA Dynamically Programmable Gate Array – Fine-grained programmable array where each processing element has a small, local configuration memory allowing processing elements to change instructions, array-wide, on a cycle-by-cycle basis.

FPGA Field Programmable Gate Array – A collection of configurable processing units embedded in a configurable interconnection network.

functional density Computations performed per unit space-time. Usually measured in $\mathrm{Ops}/\lambda^2\mathrm{s}$.

functional diversity The number of different functions which are resident and rapidly accessible from a unit of computational area. The density of instructions stored on a general-purpose computing device.

general-purpose computing Computing using devices which can be configured to solve any number of computing tasks.

iDPGA Dynamically Programmable Gate Array with input retiming registers – A DPGA including input retiming registers.

input depth The temporal range of input retiming registers in the iDPGA or similar architectures.

input folding A style for reducing the amount of active switching interconnect by sharing crossbar inputs among multiple sources.

instruction The set of bits which describe the behavior of one computational unit and its associated interconnect.

instruction context See configuration context.

instruction density See functional diversity.

instruction depth Number of instructions per compute element stored local to the compute element.

irregular computing task Task which require a large sequence of different computations and where operations are heavily data-dependent.

Kolmogorov complexity Of all programs which can be used to calculate a particular set of values, the length of the smallest such one. Ultimately, this is the least number of bits into which a piece of data can be described. Kolmogorov complexity is, primarily, a conceptual description of the lower bound as there is no algorthimic way to find such the bound. See any information theory text such as [CT91].

lambda (λ) – half the minimum feature size in a silicon process. Lambda is used to normalize out the effects of different process sizes when comparing implementations. Area normalized to λ^2 units is roughly comparable between processes which differ primarily in feature size.

low instruction entropy Computing tasks which require a limited set of operations with very regular flow, admitting to heavy compression of instruction distribution requirements.

lookup table A small, typically programmable, memory where the address bits act as inputs and data read out serves as an output. An n-input, m-output lookup table can implement any, deterministic mapping between n input bits and m output bits. We frequently refer to a k-input, 1-output lookup table as a k-LUT.

LUT Look Up Table – see lookup table.

MATRIX Multiple ALU architecture with Reconfigurable Interconnect – A flexible general-purpose computing architecture which defers binding of instructions and instruction resources until use. Instruction storage and distribution resources are unified with datapath compute, memory, and interconnect resources, allowing the basic instruction architecture to be defined at run-time.

metaconfiguration A higher and more primitive level of configuration than traditional instructions which defines the sources and distribution paths for dynamic control including instructions. See multi-level configuration.

microcycle One primitive machine cycle on architectures which evaluate logical tasks over several smaller clock cycles.

multicontext Having more than one configuration for the entire general-purpose device. Usually used to refer to devices or architectures which hold multiple such configurations on chip. Also used to describe evaluation schemes which compute a result using more than one device-wide configuration.

multi-level configuration Hierarchical configuration where higher levels of configuration describe the architecture, behavior, and distribution used by lower levers of configuration. See metaconfiguration.

output folding A style for reducing the amount of active switching interconnect by sharing crossbar outputs among multiple sinks.

partial reconfiguration The ability for individual or small numbers of processing units to change instructions without requiring an entire reload of all instructions across a general-purpose computing device.

pinst primitive instruction – the set of bits which control one compute block and its associated interconnect and state management on one cycle of operation. One can think of a typical RISC instruction as a pinst. Equivalently, the field of a VLIW instruction which controls one functional unit may be considered a pinst. This term was coined to minimize confusion relative to the term instruction [DW97].

quasistatic Changing, but on an time scale much slower than standard operation. An intermediate point of activity between dynamic and static.

quasistatic instruction distribution Instructions which change during an application, but do so slowly compared to the rate of execution. A quasistatic instruction might be in effect for hundreds of cycles before changing.

Rent's Rule An empirical relationship between the number of i/o's in and out of a cluster of logic and the number of logical elements inside the logic ($N_{io} = CN_{gates}^{p}$). See [LR71].

regular computing task Tasks which need to repeatedly perform the same collection of operations to a large amount of data with little data-dependent flow control.

retiming Changing the time at which particular events occur. In this work, used largely to describe the transportation of signals forward in time between the point in time when they are generated to the point in time when they are consumed.

robust architectural points Design points where we can bound the inefficiency to some constant percentage when the task has different characteristics from the architecture.

***RP*-space** A high-level abstraction of the reconfigurable computing design space parameterized by key instruction and interconnect features.

run-time reconfiguration The ability to change device configuration during a computational task.

segmentable datapath A SIMD controlled n-bit datapath which can be dynamically or quasistatically reconfigured to treat the datapath as k, $\frac{n}{k}$-bit words, for certain, restricted, values of k.

subarray An organizational unit in array architectures composed of multiple processing elements but not the entire device. In the DPGA and TSFPGA, the subarray defines the extent of local interconnect and the set of processing elements which share common resources such as decoders and instruction distribution.

spatial transport Movement of intermediate data in space from the point of production to the point of consumption.

static Showing little change; characterized by a lack of movement, animation or progression. In this context used primarily to distinguish values and instructions which do not change during an operational epic. Contrast with static and quasistatic.

static instruction distribution Instruction distribution where instructions are set at the beginning of a computational task and do not changed during execution.

programmable computing architectures General-purpose computing architectures which heavily and rapidly reuse a single or small number of active computing resources for many different functions (*e.g.* conventional microprocessors).

tau (τ) The delay parameter for a process. One τ is the delay required for one inverter to drive a single, equally large inverter.

temporal pipelining Reusing general-purpose resources in time to evaluate different components of a single logical task. Like spatial pipelining, the result is produced after traversing a number of pipelining stages. Unlike spatial pipelining, the same physical resources are used to evaluate each stage of the pipeline. Temporal pipelining reduces spatial requirements, whereas spatial pipelining increases throughput.

temporal transport Movement of intermediate data in time from the microcycle on which the value is produced to the one where it is consumed.

timestep A particular microcycle in the evaluation of a computing task.

time-switched input register An input register supporting data retiming on architectures which time-switch their interconnect. The input register loads the value from its associated network output only when the current timestep matches a programmed value.

TSFPGA Time-Switched Field Programmable Gate Array – Fine-grained programmable array where the physical interconnect is shared and switched in time.

yielded computational density The effective computational density which an application or task extracts from a computational device. Mismatches in datapath granularity, interconnect richness, or control may cause a device to provide computational capacity below its peak.

References

[CT91] Thomas Cover and Joy Thomas. *Elements of Information Theory*. John Wiley and Sons, Inc., New York, 1991.

[DeH96] André DeHon. Reconfigurable Architectures for General-Purpose Computing. AI Technical Report 1586, MIT Artificial Intelligence Laboratory, 545 Technology Sq., Cambridge, MA 02139, October 1996. Anonymous FTP `publications.ai.mit.edu:ai-publications/1996/AITR-1586.ps.Z`.

[DW97] André DeHon and John Wawrzynek. The Case for Reconfigurable Processors. preprint contact `amd@cs.berkeley.edu`, January 1997.

[LR71] B. S. Landman and R. L. Russo. On Pin Versus Block Relationship for Partitions of Logic Circuits. *IEEE Transactions on Computers*, 20:1469–1479, 1971.