

Euro DesignCon 2005

Evaluation of Temporal-Spatial Voltage Scaling for Processor- Like Reconfigurable Architectures

Thomas Schweizer, Julio Oliveira Filho, Tobias Oppold,
Tommy Kuhn, Wolfgang Rosenstiel
Tübingen University
Wilhelm-Schickard-Institute, Computer Engineering,
Sand 14, 72076 Tübingen, Germany
crc@informatik.uni-tuebingen.de

Abstract

Supply voltage reduction leads to quadratic savings on power/energy consumption at the cost of longer execution delays. These days, this fact is largely used to convert the task idle time into power/energy gain on general-purpose and embedded processors. However, such voltage management requires often the use of extra software layers and intelligent algorithms and therefore an additional overhead for the system in terms of power and performance. In this work, we show how spatial-temporal voltage management at the instruction level can be used to optimize power/energy consumption without performance penalty in highly reconfigurable architectures.

Author(s) Biography

Thomas Schweizer works as a research assistant in the Computer Engineering Department of the Tübingen University. There he is part of the Hardware Synthesis Group and his research topics are low power, reconfigurable systems and system level design.

Professor Wolfgang Rosenstiel received his Ph.D. in 1984 from the Karlsruhe University. Since 1990 he is Professor (Chair for Computer Engineering) at the Wilhelm-Schickard-Institute for Informatics (WSI), Tübingen University, as well as Director of FZI Department "System Design in Microelectronics". He is member of the Editorial Board of several journals. He is on the executive board of the German edacentrum. His special interests are in electronic design automation, especially synthesis, co-design, verification, and modelling, computer architecture and artificial neural networks.

1. Introduction

The continuing trend in applications for increasing functionality, performance and integration is leading to designs with bigger power dissipations and energy needs. Power dissipation implies additional packaging and cooling environment costs while energy is a major impact factor for battery lifetime. Such aspects become even more critical when considering mobile and embedded applications. Because power dissipation is directly proportional to switching frequency and quadratically related to the source voltage, these two parameters are the focus of the major power reduction techniques, independently whether the problem is considered at gate, system or any intermediary design level. Frequency management techniques try to identify performance non-critical modules or operation modes where the system may run at lower clock frequencies. Similarly, source voltage reduction also leads to power gain, but here on the cost of longer gate propagation delays which usually also demand some kind of frequency management. Both groups of techniques and their combinations are known to lead to power dissipation improvement. However, if a metric that also considers the overall system performance is used, such as Energy-Delay product [1], no meaningful gains are obtained. We point out the following reasons for why most power improvement techniques fade the performance:

- 1) Frequency and voltage management implies additional area and computational costs. In other words, such management often requires the use of extra software layers and intelligent algorithms, and therefore an additional overhead for the system in terms of power and performance.
- 2) Hardware or architectural improvements are frequently captured on the design elements that will be non-mutable after fabrication. Examples are steadily determined voltage islands or centralized power controllers. Therefore, the proposed solution lacks the necessary flexibility to attend, after fabrication, a major number of applications and their specific power needs. Up to date, even reconfigurable platforms do not offer power management resources that are themselves also reconfigurable.

In this work we propose an approach to deal with such problems based on two main pillars. First, we exploit the slack time of operations from a given application in order to reduce the voltage under which they are executed. Our approach is similar to the techniques used at gate level, where gates outside from the critical path are supplied with lower voltages. However, we consider that idea at instruction level which leads to greater flexibility after the system fabrication. Additionally, considering the voltage management at instruction level seems to be wiser as at task level in the sense that it requires less area/computational overhead. Second, we include the voltage management as a natural aspect of processor-like reconfigurable architectures. Such type of architectures allows it to instantiate and to execute within one clock cycle exactly that part of a circuit that is needed in this cycle. We extend that concept for voltage supplies, so that exactly the appropriate voltage for each processing element of the architecture is selected in few clock cycles.

We focus our work on obtaining power gains without prejudice of the overall system performance. Therefore, we are concerned on metrics that consider simultaneously the

power/energy gains and the performance of the system. Such goal is achieved by (a) diminishing the overhead for voltage and frequency management to a minimum and by (b) providing higher flexibility on which modules and when the power management takes place. For the first objective, our proposal eliminates any need for clock management through operational systems or hard power controllers by determining how it is done at compile time. For our second objective, we exploit the coarse-grained array structure of the architecture and the fast reconfiguration mechanism in order to provide an individual supply voltage selection for each processing element. Hence it is possible to scale spatially and temporally the voltage on the system. The supply voltage selection is incorporated in a natural way on processor-like reconfigurable architectures, so that the necessary additional hardware may also be kept to a minimum.

The remainder of this paper is organized as follows: in the next section, we discuss and classify voltage scaling methodologies according to their spatial and temporal aspects. Section 3 discusses the necessary architectural features to support temporal-spatial voltage scaling, followed by the methodology we used to synthesize and measure power consumption. Section 4 explains in details the voltage scaling methodology proposed on this paper and contrasts it to pure spatial and pure temporal approaches. A small example is depicted where no damage to the performance is achieved only if temporal-spatial voltage scaling capability is provided. We present our experimental results on Section 5. Section 6 concludes the paper and foresees our future work.

2. Classification of Power Optimization Methods

In order to clarify the concept of temporal-spatial voltage scaling developed throughout this work, it is convenient to group the actual existing power optimization methods under its temporal and spatial characteristics as well as under the granularity of the digital circuit components where they are employed. We join both aspects through the diagram presented in Figure 1.

On the right-most arrow we group the approaches that target power optimization exclusively by creating bounded regions for different voltages (voltage islands) and/or where for each component of the system a suitable working voltage is determined. By saying exclusively, we stress the fact that such approaches determine the spatial distribution of different voltages before production time, therefore, with no possibility of scaling them after fabrication. We say that, on that frontier, the power optimization problem is considered exclusively on the spatial aspect of the voltage distribution. In counterpart, on the left-most arrow we group the approaches that target power optimization solely by dynamically scaling the voltage of the complete system. In these cases one unique voltage, scalable at running time, is applied for all the components. On that frontier, the power optimization problem is considered exclusively through the temporal adjustment of the supply voltage.

Two distinct intermediary groups appear between those extremes and are organized from right to left according to when the voltage scaling strategy is defined. The techniques that determine how the different voltages are distributed (voltage islands) among the system components after the fabrication but not at running time are grouped on the inner-right

arrow designated as configurable. If such voltage spatial distribution strategy is defined at running time we say the technique explores voltage reconfiguration. Such approaches are grouped on the inner-left arrow designated as reconfigurable. Both intermediary groups, namely configurable and reconfigurable, comprehend our concept for temporal-spatial voltage scaling.

Additionally, the approaches differentiate, from bottom to top, according to which system granularity is considered for the voltage scaling. From bottom to top, we distinguish several ranks, namely transistor, gate, logic blocks, processing elements, SoC and PCB components. At each rank it is possible to consider, at least theoretically, full spatial, full temporal or temporal-spatial strategies to realize the voltage scaling.

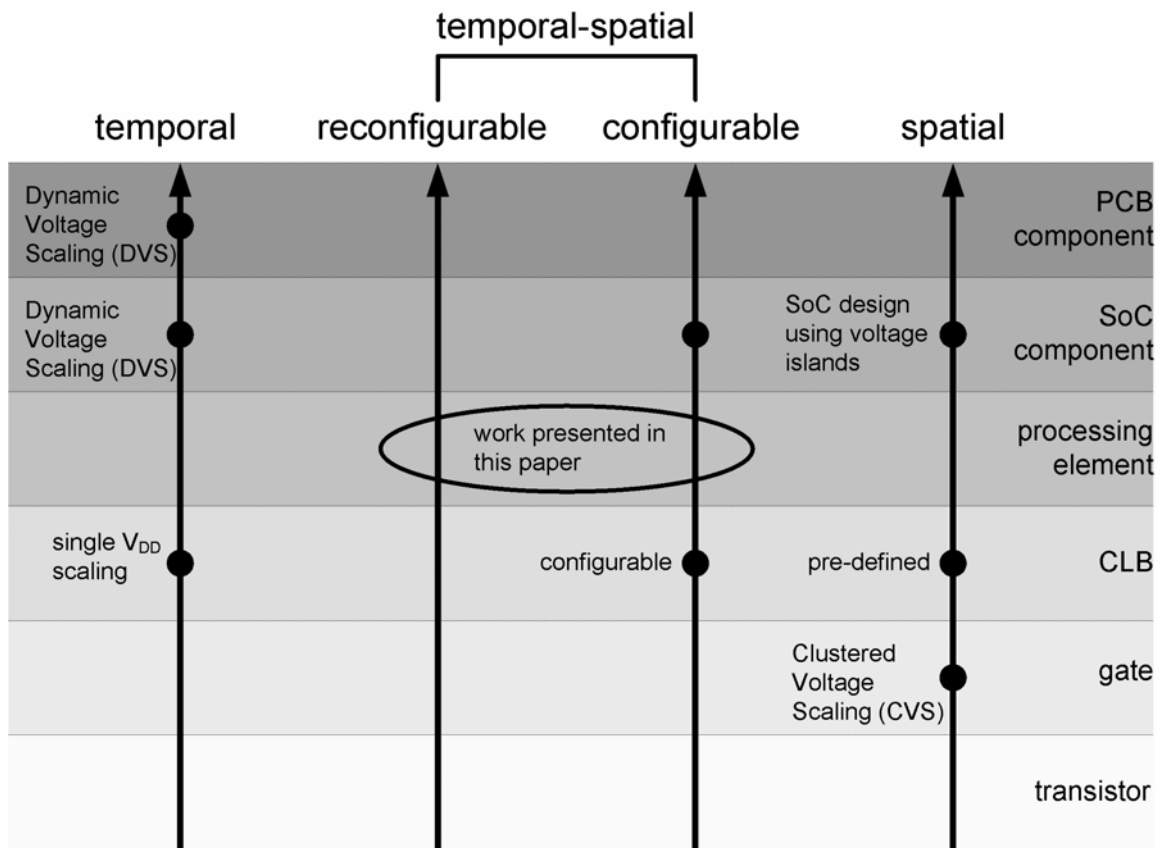


Figure 1: Classification of different voltage scaling technique.

Several examples of power optimization approaches which are presented in the literature and influence the present work are shown in the diagram. Dynamic Voltage Scaling [2] is proposed as a technique to dynamically adapt the supply voltage of a microprocessor.

Spatial distribution of the elements working voltage is predefined before fabrication in two well known works. First, Clustered Voltage Scaling technique [3] is used to let the gates from non-critical paths run at lower voltage. The technique fits well for ASIC design, however lacks of flexibility. Second, [4] describes a technique to determine voltage islands on SoC designs. Once again, the technique considers that no temporal voltage scaling will take place, and therefore is grouped on the fully spatial side. Fei Li showed [5] that on FPGAs configurable dual-Vdd reduces power significantly compared to single Vdd scaling or pre-defined Vdd levels.

Our work is concentrated, as indicated on the diagram, on temporal-spatial voltage scaling, and therefore provides the necessary flexibility required for reconfigurable systems. Moreover, we apply the voltage scaling on functional units and processing elements, which allow a reasonable faster (re)configuration of the used supply voltage in comparison to techniques that make it at system or PCB level. Simultaneously, when considering that granularity level, it is possible to expunge costs relative to operational system management and frequency scaling. In other words, our technique provides power improvement without degrading the overall system performance.

3. Target Architecture

3.1. Description

Before introducing our temporal-spatial voltage scaling technique, we consider some aspects and necessary modifications of the underlying hardware architecture. As an initial point, we consider the processor-like architecture proposed in [6], that consists of uniform processing elements (PEs, see Figure 2) that are connected by a nearest neighbor interconnection network. Each PE features a configurable finite state machine (FSM), a two-input-port functional unit (FU), which comprises data and status signals, a register set, and multiplexers used to implement communication with the interconnect network. The context memory stores, for each context, the information of which internal connections, registers and multiplexers are to be used. As well, it determines the data/status input sources, the output ports and registers and the function to be performed on the FU. The FSM selects an entry of the context memory depending on the current state. The state transitions of the FSM are controlled by the status signals. The initial focus of this work is based on the processing elements and its internal structure. However, we foresee further improvement of our technique for interconnection network and memory modules on the future work section.

In order to provide temporal-spatial voltage scaling capabilities on the architecture, while maintaining the voltage scaling delay as fast as possible, the architecture should be expanded (see Figure 2). In that sense, the context memory is extended in order to include configuration bits that will select the voltage used to supply the functional unit at a given context. Such extension has a minor impact on the size of the context memory as it requires $n = \log_2 V$ bits in each memory line, where V is the number of available voltages. A specific Voltage Switching mechanism is necessary to diminish the voltage swing after a new voltage is selected and thus the necessary delay to stabilize the functional unit. Level converters are necessary only on the output of the functional unit

in order to drive the signals back to the higher voltage level. Input ports need no level converters for the voltages used on that work.

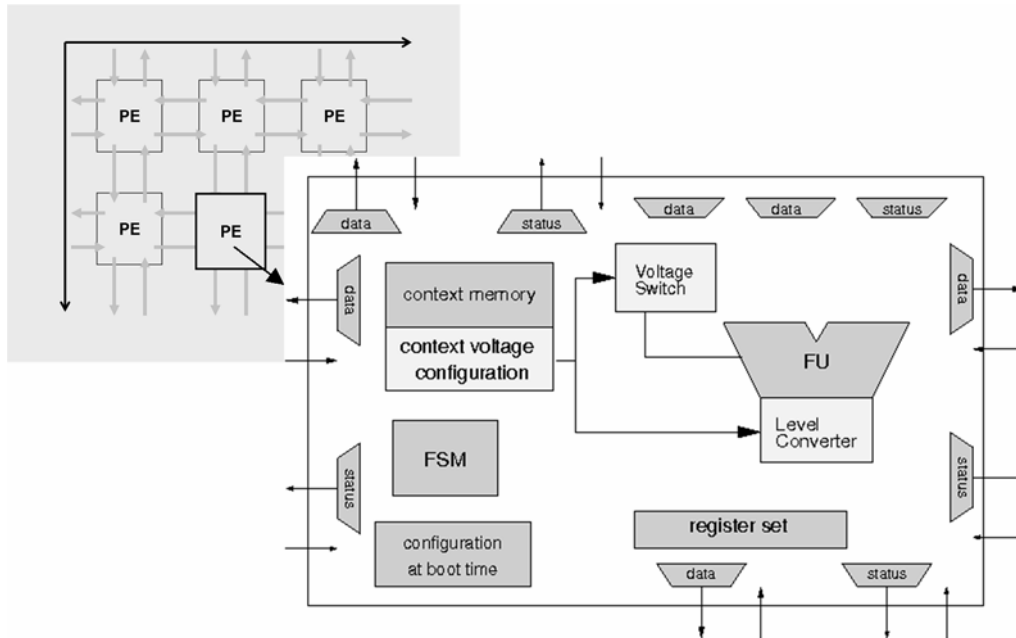


Figure 2: Extension of the processing element.

3.2. Delay and Power Evaluation

As described above we want to realize the temporal-spatial voltage scaling by the possibility of changing the FU's voltage during runtime. In order to be able to assess the benefit of temporal-spatial voltage scaling for processor-like reconfigurable architectures we synthesized the FU at different voltages with a commercial synthesis tool and carried out a power estimation at RT level based on the default switching activity parameters from the power analysis tool. We synthesized the FU with 1.0 V and 1.2 V. All components of the FU are 32 bit wide. We used the Virtual Silicon's PowerSaver standard cells as technology-library, whose building blocks are characterized at the mentioned voltage levels, to accommodate voltage scaling. The synthesis tool was configured to guarantee that at different voltages the same synthesized circuit would be obtained. A summary of the dynamic delay and the power results are shown in **Fehler! Verweisquelle konnte nicht gefunden werden.** Table 1 and Table 2, considering a clock period of 4.5 ns and a 130 nm process.

	1.2 V	1.0 V
* 16x16	4.06 ns	5.45 ns
+	2.68 ns	3.48 ns
-	2.69 ns	3.49 ns
shift	1.38 ns	2.29 ns
==	2.12 ns	2.80 ns
!=	2.10 ns	2.80 ns
>	2.30 ns	3.08 ns
>=	2.43 ns	3.15 ns
<	2.21 ns	3.04 ns
<=	2.36 ns	3.14 ns

Table 1: Delay of the FU components at different voltage levels.

	1.2 V	1.0 V
Power FU	3.71 mW	2.32 mW

Table 2: Power consumption at different voltage levels.

4. Temporal-Spatial Voltage Scaling

The synthesis results explained on Section 3 describe the power consumption for one functional unit as a function of one unique parameter: the supply voltage, whereas two parameters are considered in order to calculate the delay of the synthesized components: supply voltage and which operation is to be executed. For a given timing constraint, and if no operator chaining is employed, the assignment of the operating voltage to operations achieves maximum power reduction when every functional unit is executed on the lower voltage at which it does not violate the timing constraint. That voltage assignment strategy is a direct consequence of the fact that the slack-time of components is used in order to diminish their power/energy consumption. Although the best power consumption configuration scheme is achieved through the method mentioned above, only when temporal-spatial voltage scaling is provided it is possible to optimize area and/or performance completely independently from the chosen voltage assignment. In other words, scheduling and binding phases may be executed unaware of the voltage scaling.

In the following section, we show the conclusion above through an example. We map the fragment of a data-flow graph (Fig 3.a) onto two functional units under different voltage scaling methods, so that power consumption is optimized. We deliberately set our time constraint to 4.5 ns which implies, according to the delays in Fig 3.b, that all

multipliers have to be assigned to 1.2 V. In counterpart, the subtractor may be assigned to a lower voltage, as its low voltage delay does not violate the timing constraint.

For the first scenario, we employ a totally spatial voltage scaling mechanism, as depicted in Fig 3.c. The two available processing elements are distributed into two fixed voltage areas, which do not change that characteristic over time. As multipliers cannot be binded on the lower voltage area, and there is only one processing element per voltage area, three clock cycles are necessary because the two initial multipliers should be executed in subsequent cycles. Note that information about the voltage assignment is considered in both scheduling and binding phases. Total consumed energy is $605 \cdot 10^{-13}$ J and energy-delay product is $817.08 \cdot 10^{-21}$ Js.

In case a completely temporal solution is used (Fig. 3.d), all the processing elements have a unique supply voltage level for a given voltage clock cycle. Although that voltage may vary over time, such variation will affect all the elements. That allows the two initial multiplications to be carried out together, but that speedup advantage disappears because the next multiplier may not execute on the same clock cycle as the subtractor without violating the time or prejudicing the power optimization. Once more, voltage scaling is to be considered on the scheduling and leads to performance decrease. Also for that scenario the calculated energy and energy-delay are similar to the totally spatial solution.

On our proposed temporal-spatial voltage scaling methodology, each functional unit may be reconfigured independently from spatial position or time to a desired voltage. For such approach, both initial multipliers may run together on the first clock cycle as well as the last multiplier is executed in parallel with the subtractor on a power optimal configuration scheme (Fig. 3.e). Note that the actual voltage assignment does not interfere on scheduling (temporal) or binding (spatial) aspects. As a consequence, though the energy consumption for this case is still the same we have $544.72 \cdot 10^{-21}$ Js as energy-delay product, that is 33 % less. That improvement on the energy-delay product may be interpreted as increased power efficiency, or in other words a better relation between energy consumption and performance.

By providing fast (~clock cycle) and independent (each PE) voltage reconfiguration, which characterize temporal-spatial voltage scaling, the energy optimal solution is met independently from the scheduling/binding phase, and vice-versa, performance and area are not influenced by use of the voltage scaling technique and turns out to be completely determined exclusively by the natural data-dependency and number of available functional units.

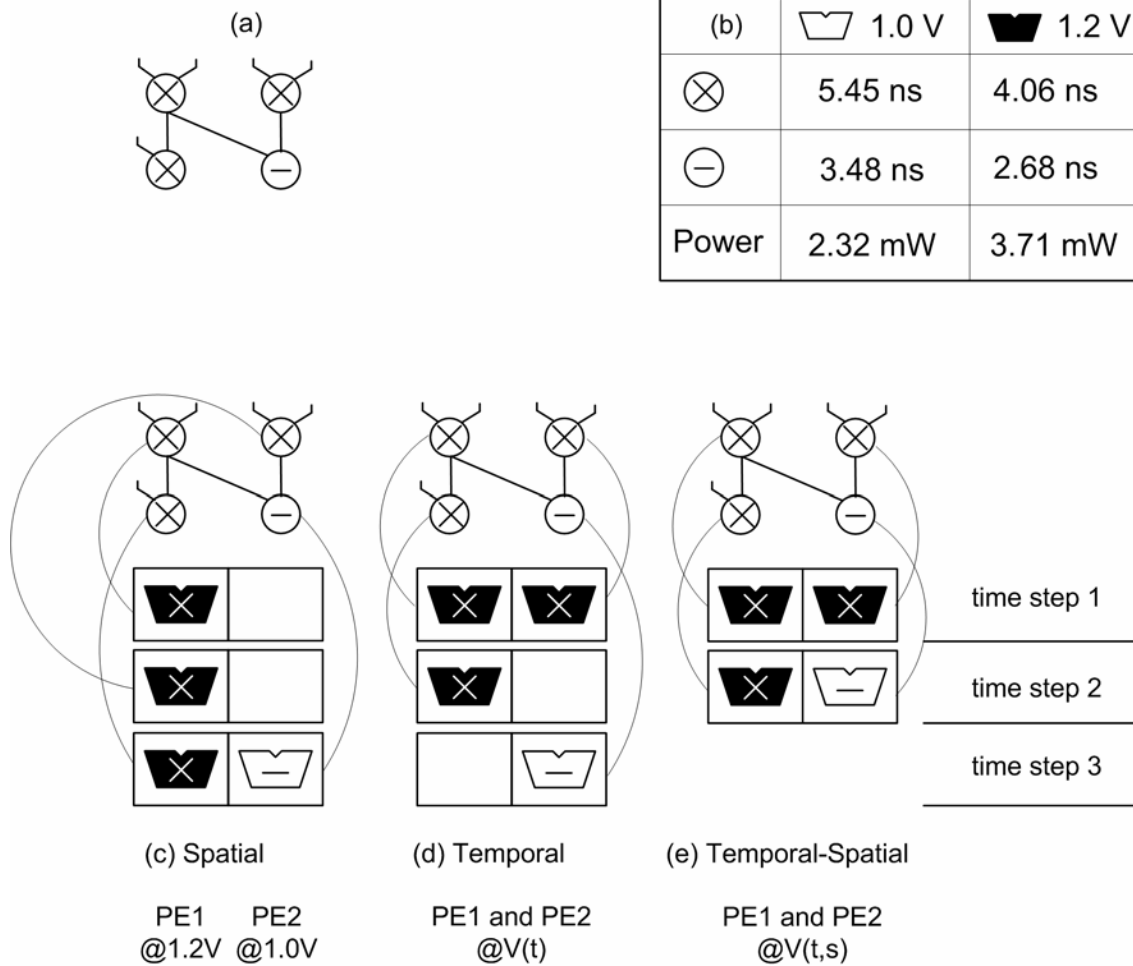


Figure 3: Voltage scaling methods.

If operator chaining is used, some other aspects should be considered for ensuring independency between voltage assignment and scheduling and binding phases. Such considerations are to be investigated on our near future work and will not be described here.

5. Experimental Results

We applied the above voltage assignment method to the ray casting [7] algorithm. Ray casting is a real-life application for the visualization of 3D scientific and medical data. The ray casting algorithm is usually implemented in a pipeline consisting of five major stages namely *ray setup*, *ray traversal* and *voxel fetch*, *resampling*, *classification* and *shading*, and *compositing*.

For our evaluation we have chosen an architecture [8] optimized for low area requirements and we considered the stages *voxel fetch* and *resampling*. The proposed architecture uses 4 processing elements. The *trilinear interpolation*, part of the

consumption at the highest voltage. Obviously, it is achieved only if all operations may be assigned to the lowest available voltage. The maximum achievable energy consumption gain for the above described experiment is 37 %. The real achieved gain is obtained by multiplying this value by the rate O^{ps}/O , where O^{ps} is the number of operations able to be assigned to the lowest voltage and O is the total number of operations. For our example, a real energy consumption gain of 27.7 % is achieved.

For the Voxel Fetch algorithm mapping, 15 out of 17 operations (see Figure 5) may be executed at 1.0V. That leads to a total energy reduction as well as an energy-delay product reduction of 32.7%.

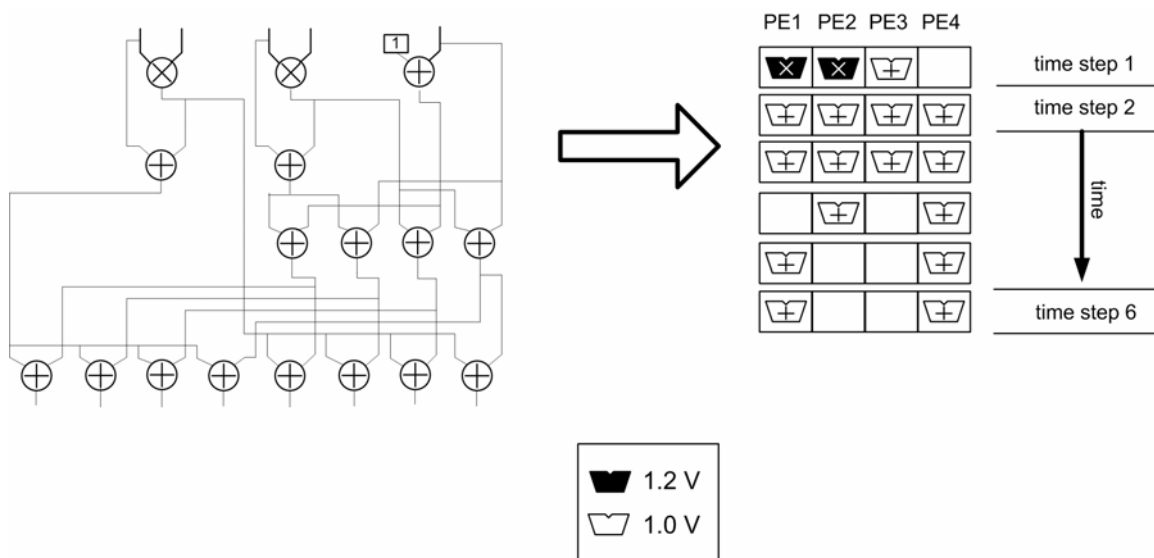


Figure 5: Mapping and voltage assignment of voxel fetch.

6. Conclusions and Further Work

We proposed a voltage scaling technique for processor-like reconfigurable hardware which satisfies the timing constraints without frequency scaling. The power dissipation in each time step, the total energy consumption as well as the energy-delay product can be reduced enormously by temporal-spatial voltage assignment. By providing fast and independent voltage reconfiguration, the power optimal solution is chosen independently from the scheduling/binding phase.

We described the hardware extension of the processing element from the CRC-Model to implement the voltage reconfiguration mechanism. In a next step a quantitative study with regard to the additional reconfiguration time needs to be carried out to justify the programmable voltage switch.

7. Acknowledgements

This work is funded by DFG under RO-1030/13 within the ‘DFG Priority Program 1148’ which is focused on reconfigurable computing systems.

8. References

- [1] Ricardo Gonzalez and Mark Horowitz. Energy Dissipation in General Purpose Microprocessors. *IEEE Journal of Solid-State Circuits*, 1996.
- [2] Trevor Pering, Thomas Burd, and Robert Brodersen. Voltage Scheduling in the lpARM Microprocessor System. In *Proceedings of the International Symposium on Low-Power Electronics and Design ISLPED'00*, 2000.
- [3] Kimiyoshi Usami and Mark Horowitz. Clustered Voltage Scaling Technique for low-Power Design. In *Proceedings of the 1995 international symposium on Low power design*, 1995
- [4] David E. Lackey, Paul S. Zuchowski, Thomas R. Bednar, Douglas W. Stout, Scout W. Gould, John M. Cohn. Managing Power and Performance for System-on-Chip Designs using Voltage Islands. In *Proceedings of the 2002 IEEE/ACM international conference on Computer-aided design* , 2002
- [5] Fei Li, Yan Lin and Lei He. FPGA Power Reduction Using Configurable Dual-Vdd. In *Proceedings of the 41st annual conference on Design automation*, 2004
- [6] T. Oppold, T. Schweizer, T. Kuhn, W. Rosenstiel. Cost Functions for the Design of Dynamically Reconfigurable Processor Architectures. *Workshop on Synthesis And System Integration of Mixed Information technologies (SASIMI) 2004*, Kanazawa, Japan.
- [7] R.A. Drebin, L. Carpenter, and P. Hanrahan. Volume Rendering. *Computer Graphics*, vol. 22, no. 4, 1988
- [8] T. Oppold, T. Schweizer, T. Kuhn, W. Rosenstiel, U. Kanus, W. Strasser. Evaluation of Ray Casting on Processor-Like Reconfigurable Architectures. In *Proceedings of the 2005 International Conference on Field Programmable Logic and Applications (FPL) 2005*, Tampere, Finland.