

Accelerate Low-Power Design with PICO Extreme Power

1. Introduction

The expansion of the consumer market for mobile devices as well as thermal considerations in plugged-in devices has placed power as one of the key design concerns. At the same time power has gained in importance, traditional constraints in terms of performance and design time remain as crucial as ever to the success of a system. This convergence of strict design challenges is pushing designers to rethink IP creation and to find new methodologies to achieve design goals. Designers who select the right tools and methodologies to master these challenges have the winning combination to complete IP design on time, on budget, and on target to satisfy specification requirements.

Traditionally, power optimization has been considered as a final step in RTL generation to be attempted once both performance and area requirements have been met. There are several problems with this approach:

1. Once RTL is functionally complete and in verification, power optimization can be a complex and risky approach that can delay tape-out.
2. While the computational capabilities of mobile devices have grown exponentially, battery life and capacity has only improved at an almost linear rate. Opportunities to change the power profile of an algorithm once RTL is complete are very limited.
3. It is too late in the design process to consider advanced optimization techniques such as clock gating. This can leave a design burning up to 50% more power than necessary to complete the functionality.

Given the growing importance of power optimization and the problems mentioned above, designers have to rethink their design methodology. The biggest power optimizations opportunities are found at the system and architecture level, and then at the implementation level (see Figure 1). At the same time, a new generation of tools makes it possible to analyze an architecture and its impact on the power consumption in minutes and hours rather than days enabling fast design iterations to converge on a good solution. The time and effort required to generate correct RTL in traditional flows do not allow for architectural exploration. The designer is limited to pre-implementation analysis for any form of guidance towards power reduction.

PICO Extreme Power is the industry's first algorithmic synthesis tool that automatically optimizes the power consumption at the system and architecture level using a variety of techniques including automatic multi-level clock gating insertion along with the necessary control logic. PICO Extreme Power has delivered savings of up to 50% using

this technique. This white paper presents an introduction to algorithmic synthesis and how PICO Extreme Power addresses the requirements of low-power design.

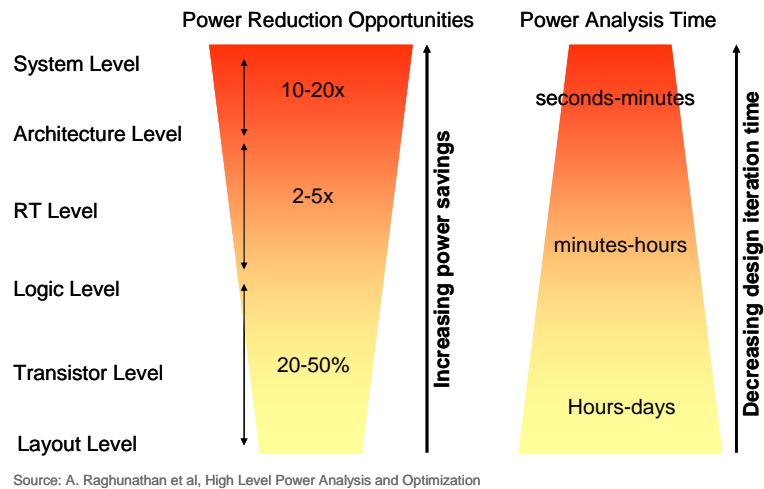


Figure 1: System and architecture level provide the biggest power optimizations opportunities

2. Algorithmic Synthesis

Algorithmic Synthesis (AS) refers to a class of hardware design tools, which raise the level of abstraction for design capture from RTL to a programmatic language such as C. One of the key advantages of AS tools is that efficient hardware implementations are derived from untimed, sequential C algorithms. This allows the designer to focus on the algorithm and be shielded from the error-prone steps involved in writing/verifying RTL.

The applicability of AS tools like PICO Extreme Power is primarily in the domain of application accelerator. Figure 1 shows a typical block diagram for a contemporary consumer IC.

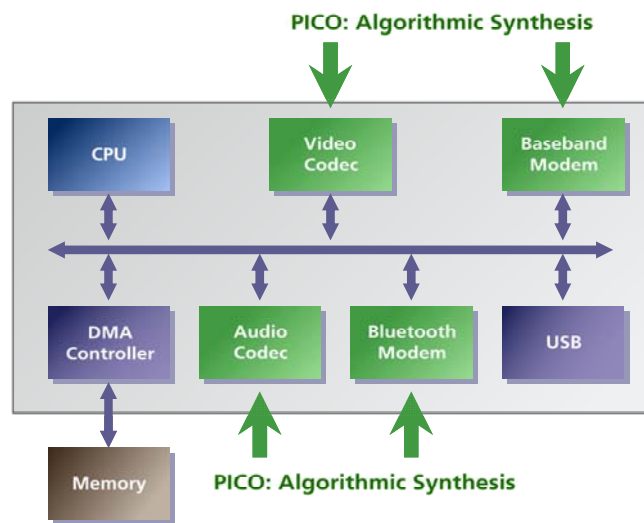


Figure 2: Consumer IC Block Diagram

The IC is comprised of two types of IP: standard IP and application accelerators. The standard IP covers elements such as processors, memory controllers, bus structures, which are easily available as off-the-shelf components. They are functionally necessary, but add little in terms of differentiation with a competing IC. Application accelerators shown in green are the ones that provide product differentiation. Application accelerators deliver the value in the solution, but are also the most difficult elements in the IC to design and verify. PICO Extreme Power allows the designer to optimize the power consumption of his design as well as explore different architectural options and different implementations.

3. Multi-level Clock Gating

Clock gating is a popular technique for reducing dynamic power consumption at the RTL level. The basic premise of this technique is that portions of a computational datapath can be turned on and off depending on dynamic processing requirements by shutting off sections of the clock tree network.

In traditional RTL design methodologies, inserting clock gating at a block level is usually a time-consuming manual effort because it requires the knowledge of when the block is inactive. Another problem with the manual approach is in the verification of the final hardware. If blocks are created in isolation, it is very easy to mistakenly clock gate a datapath in one block which affects another block in the IP. In many cases, an error in clock gating in one part of the IP can lead to a deadlock in a completely different part of the IP. Since the deadlock will appear during runtime when clock gating is active, it can be difficult to debug.

PICO Extreme Power introduces a major innovation in algorithmic synthesis -- automatic multi-level clock gating insertion -- to enable power optimizations at the system level and eliminate all manual work. Using PICO Extreme Power, the designer uses directives to specify where to insert clock gating, and PICO does the rest automatically. In all cases the user can make changes without having to impact the algorithm or the code. The key capabilities include:

1. **Coarse-grain clock gating:** PICO Extreme Power builds the clock gating infrastructure to turn off complete blocks at the top level of the design; for example, to turn off the complete quantize stage of an imaging pipeline. Clock gating is integrated directly into the PICO architectural template as shown in Figure 3. Of critical value is the control logic that will indicate when the block can be turned off. PICO Extreme Power analyzes the input code and knows all the data/control dependencies between blocks and when datapaths can be safely turned off. The generation of all clock gating enable signals is transparent to the user, and there is no need for time consuming manual analysis to decide “when” a block can be turned off. The only thing a designer has to worry about is correctly instantiating the clock gating cell specific to their fabrication process.
2. **Fine-grain clock gating:** There may be significant power saving by turning off only portions of a block. Earlier, PICO Extreme introduced an innovative

technique called Tightly Coupled Accelerator Block (TCAB) for multi-level hierarchical designs. PICO Extreme Power allows clock gating of TCABs used in a top level block or in another TCAB. An example of TCAB clock gating is the selective activation of a long latency divider within a block. Like coarse-grain clock gating, PICO Extreme Power automates clock gating insertion for TCABs hierarchically.

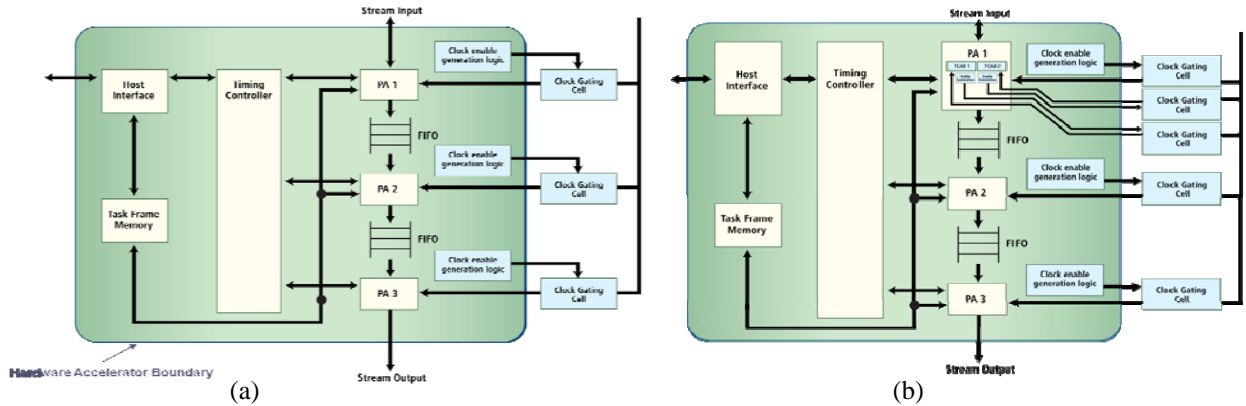


Figure 3: (a) Coarse-grain clock gating (b) Fine-grain clock gating

- Automatic functional verification:** PICO Extreme Power provides automatic functional verification to check the sequencing of clock gating for both coarse-grained as well as fine-grained clock gating.
- Integration with downstream tools:** PICO Extreme Power automatically generates waveforms in VCD/FSDB formats to enable power measurement in down-stream power analysis tools.

As a result of these capabilities, PICO Extreme Power allows designers to retain all the productivity benefits of automated synthesis and verification, including reduced design and verification time and the ability to react very rapidly to changes in the design specification, while optimizing the IC power consumption.

The following examples demonstrate the benefits of clock gating.

- Table 1 shows the benefits of clock gating for a wireless design -- a cellular receiver chain. The combination of coarse- and fine-grain clock gating reduces the dynamic power consumption by more than 50%. This level of power saving comes from PICO Extreme Power exploiting the flexibility provided by the multiple operating modes of the IP.

Design	Power Consumption	Power Savings
No clock gating	4.81 mW	
Fine-grain clock gating	2.43 mW	50 %
Coarse- and fine-grain clock gating	2.29 mW	53 %

Table 1: Power reduction using clock gating in a wireless design

2. Table 2 shows that clock gating reduces the power consumption by 22% for an HD video scaler design.

Design	Power Consumption	Power Savings
No clock gating	22.5 mW	
Fine-grain clock gating	21.8 mW	4.73 %
Coarse- and fine-grain clock gating	19.0 mW	22.4 %

Table 2: Power reduction using clock gating in HD video scaler design

3. The design for a low density parity check (LDPC) decoder for the next generation wireless handset SoC achieved 23.5% reduction in dynamic power over an identical design using a standard flow.
4. An evaluation of the effectiveness of the approach using 8 complex applications in video, imaging and wireless domains demonstrated the following:
 - Up to 50% reduction in dynamic power for executing a single task and up to 30% savings while executing a large number of tasks
 - Average power reduction of 22% for a single task and 15% over multiple tasks

As demonstrated by these examples, the multi-level clock gating capability in PICO Extreme Power provides significant power savings – more than 50% for some applications. These savings are over-and-above what can be achieved with gate-level clock gating in down-stream tools. The process is fully automated and easy to use, and it eliminates time-consuming manual effort to insert clock gating, its verification and power measurement with down-stream tools.

3. Architecture Exploration for Low Power Design

Architectural exploration allows the designer to carry out what if scenario analysis. In approaching a hand-crafted design, the architect can make power trade-offs only on paper, because the effort associated with implementing multiple design approaches is prohibitively high. PICO Extreme Power makes it possible to explore multiple design alternatives, either from the same code or from alternative coding styles designed to create different hardware architectures. These alternatives can be passed through power analysis tools such as RTL power estimation to quickly evaluate the best choices. The types of tradeoff that can be made include the following:

- Design partitioning: Putting different parts of the algorithm in different blocks
- Memory architecture
- Inter-block communication architecture
- Rate-matching: Running each part of the design at an optimal rate for power minimization
- Resource-sharing

As an example, one of the scenarios which is virtually impossible to do in manual RTL is the side by side comparison of different memory architectures and different memory data

shapes. When it comes to memory architectures, there are lots of possible choices. Should there be a single bank or multiple banks? What is the number of ports per bank? All these decisions have a significant impact on the power consumption. Due to the large range of choices, a lot of time is spent upfront in the design cycle thinking about memory architecture. The reason why this can't be studied at the RTL level in a traditional flow is that memory architecture can potentially impact a significantly large part of the design. At the RTL level, a change in memory architecture can set a project back for months in implementation and verification.

Another architectural component which is unfeasible to experiment with at the RTL level is block to block communication structures. How blocks in an IP communicate has an impact on the pipelining, resource allocation, and execution rates of blocks, all of which affect the power consumption in a significant way. In some cases, changing how blocks communicate requires a similar amount of effort at the RTL level as starting from scratch.

PICO tools simplify this process by taking care of the scheduling, block rate matching and resource allocation required in a functionally correct RTL implementation. For the designer, this means that the implementation spec does not have to be perfect the first time around. Using a tool like PICO Extreme Power, frees up time in the development schedule to test out the performance and power profile of different architectural choices. Armed with the multiple implementation created by PICO Extreme Power, the designer can do a side by side comparison to select the optimal implementation which satisfies all design requirements.

To illustrate some of these concepts, Figure 4 shows two different architectures for a simple video design.

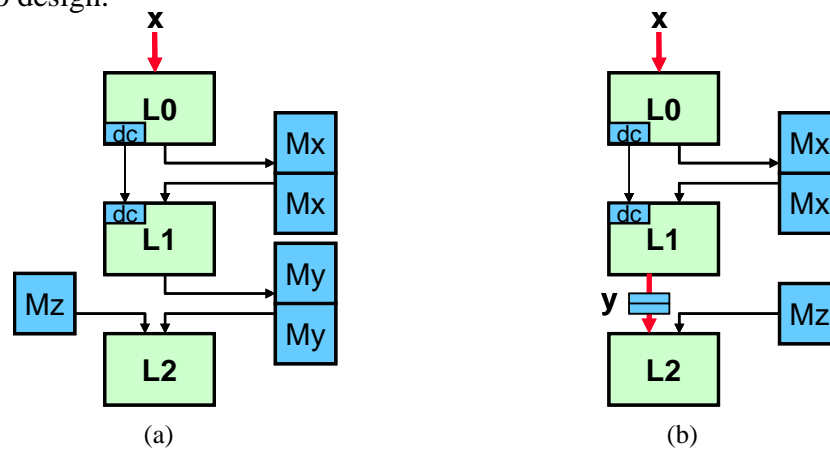


Figure 4: Two different choices for inter-block communication (a) L1-L2 communication uses a memory (b) L1-L2 communication uses a FIFO

In this design, the block L1 computes the elements of an array y , which are then used by the block L2. A simple way to pass these data elements from L1 to L2 is to use a memory "My" as shown in Figure 3(a). L1 writes the elements of y in this memory and L2 then reads them from this memory. An important aspect of this design is that the order in which elements are produced by the block L1 is the same order in which they are

consumed by the block L2. Thus, it is possible to communicate the array elements between these blocks using a FIFO (or a “stream”) as shown in Figure 3(b).

Stream-based communication has an advantage over memory-based communication in that it allows *overlapped execution* of the two blocks leading to higher performance designs. Both L1 and L2 can be executing at the same time – while L2 is using an array element, L1 is computing the next element to send to L2. However, it is not at all obvious which of these architectures is better from a power perspective.

PICO Extreme Power makes it possible to create and analyze both these architectures in a short period of time and then choose the one that minimizes power consumption. For the design described above, the area and power characteristics of these two architectures are shown in Figure 4. There were two options considered for implementing memory: as registers or as a SRAM. The area and power consumption of the SRAM was obtained from datasheets provide by a memory vendor. From the data in Figure 5, it is clear that stream-based communication is the winning choice for this design, since it is both area and power efficient.

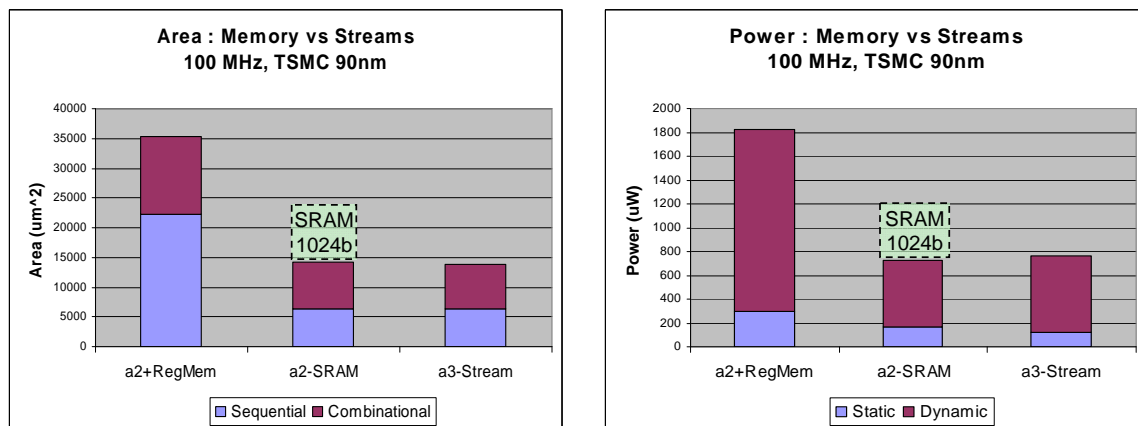


Figure 5: Area and power characteristics of memory and stream based architectures

A side by side comparison of architectural choices with data coming from real RTL implementation helps ensure that the best architecture for a given algorithm makes it to silicon on the first time. In a traditional manual design flow, if the architecture is wrong, there are only 2 choices for the problem resolution. Either delay the launch of the product and risk losing market share, or re-spin the corrected design and loose in profit margins. For these reasons, traditional handwritten RTL is not a cost effective methodology for developing IP with tight power, area and performance constraints.

4. Frequency and Voltage Scaling for Low Power Design

Another technique which can be used for achieving low power consumption is to modify the operating frequency and voltage. While these techniques can provide significant power savings, the power savings can drop the performance of the design if the RTL is not redesigned accordingly. To account for the performance degradation due to the

lowering of clock frequency or operating voltage, the level of parallelism in the design has to be increased to maintain the same level of computational performance.

PICO Extreme Power accepts the overall computational throughput as a design constraint which is independent of clock and voltage level. This allows the designer to automatically create designs optimized for several frequency/voltage points. From the resulting implementations, the designer can select the best design which exhibits the appropriate power, area and performance trade-off to satisfy the application requirements.

As an example of frequency exploration to find a low power design, Figure 6 shows power and area characteristics of four different implementations of a wireless design with a specified throughput but running at different frequencies. To keep the throughput constant, lower frequencies require design to operate at a higher parallelism. The 1x, 2x, 4x and 8x in the diagram show how much parallelism is built into the hardware.

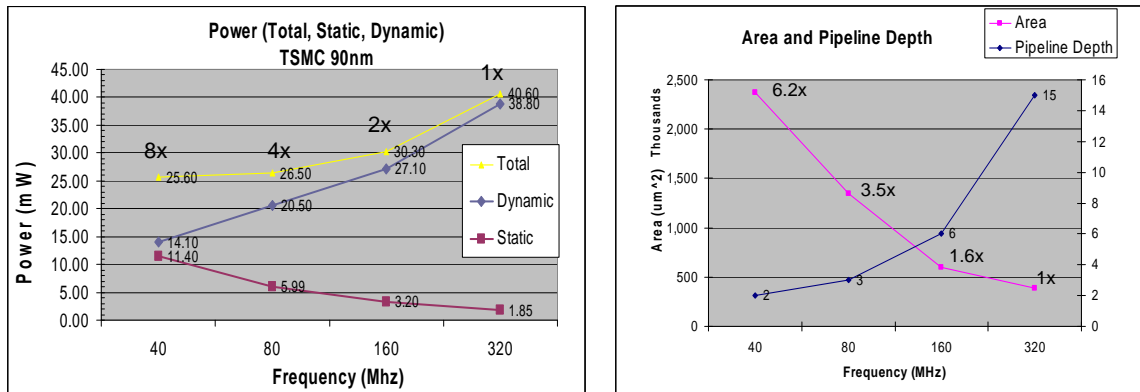


Figure 6: Exploring multiple implementations of a wireless design for different frequencies

The 40 MHz design has ~37% less total (dynamic + static) power consumption than the 320 MHz design. Since power consumption is linearly dependent on frequency and area, it is not at all obvious why that is the case. The area graph shows that as the frequency reduced, the pipelines become shorter, savings on register and logic area. Area goes up but not linearly – the 40 MHz design is not 8 times as big as the 320 MHz design.

Using PICO, it took less than a day to implement all these four designs – something, that would have been completely impractical if these designs were done manually.

6. Conclusion

Creating a low-power design involves the combination of architecture exploration, frequency and voltage scaling and advanced clock gating capabilities. Traditional handwritten RTL methodologies fail to deliver these kinds of designs in a competitive time-frame because of the design and verification complexities introduced by low-power constraints. PICO Extreme Power extends the benefits of algorithmic synthesis to meet the requirements of low-power design. The automatic generation of RTL from untimed sequential C code and integrated verification in PICO Extreme Power accelerates and simplifies the creation of low-power algorithmic IP.