

Architectural Tradeoffs for Low Power

Vojin G. Oklobdzija

Integration

Berkeley, CA 94708

<http://www.integr.com>

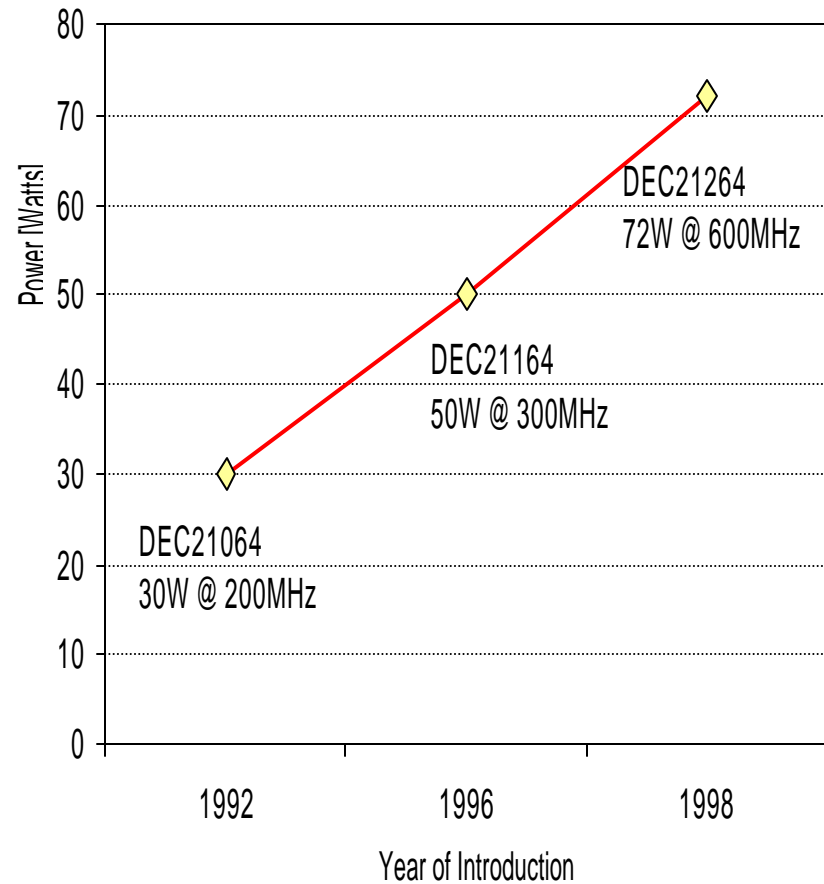
*Electrical Engineering Department

University of California, Davis, CA 95616

<http://www.ece.ucdavis.edu/acsel>

Current Prospects:

- Demand for performance is leaving an imbalance in power dissipation:
- Fabrication of a suitable device for the year 2030 will be possible (Iwai, Toshiba), but not power management.
- Power is growing approximately at 10 Watts per year.
- Power growth is threatening to limit the performance of the future microprocessors.
- *The “CMOS ULSIs are facing a power dissipation crisis”*



Current Situation:

- Frequency of operation has been doubling for each new generation
- Switching energy CV^2 has been improving at the rate of 0.5 times per generation.
- The power factor $P = CV^2f$ remained constant ($0.5 \times 2 = 1.0$).

The increase in complexity of the VLSI circuits goes largely uncompensated as far as power is concerned.

Current Situation:

- The number of transistor has been tripling for every generation.

Therefore, the expected processor performance increase is 6 times per generation (two times due to the doubling of processor frequency and the three times for the increase in the number of transistors).

- The fact that the performance has been increasing only four times per generation instead of six is a strong indications that the transistors are not efficiently used.

That means that the added architectural features are at the point of diminishing returns.

Diminishing trend in transistor utilization

Table 1. is comparing a transition from a single issue machine to a super-scalar on the example of IBM PowerPC architecture.

| Feature | 601+ | 604 | 620 | Difference |
|--------------------------|-------------|----------------|--------------------|-------------------|
| <i>Frequency</i> | 100MHz | 100MHz | 133MHz (100MHz) | same |
| <i>CMOS Process</i> | .5u 5-metal | .5u 4-metal | 0.5u 4-metal | ~same |
| <i>Cache Total</i> | 32KB Cache | 16K+16K Cache | 64K | ~same |
| <i>Load/Store Unit</i> | No | Yes | Yes | |
| <i>Dual Integer Unit</i> | No | Yes | Yes | |
| <i>Register Renaming</i> | No | Yes | Yes | |
| <i>Peak Issue</i> | 2 + Branch | 4 Instructions | 4 Instructions | ~double |
| <i>Transistors</i> | 2.8 Million | 3.6 Million | 6.9 Million | +30% /+146% |
| <i>SPECint92</i> | 105 | 160 | 225 (169) | +50% /+61% |
| <i>SPECfp02</i> | 125 | 165 | 300 (225) | +30% /+80% |
| <i>Power</i> | 4W | 13W | 30W (22.5W) | +225%/+463% |
| <i>Spec/Watt</i> | 26.5/31.2 | 12.3/12.7 | 7.5/10 | -115%/-252% |

Table 2. Transition from single issue MIPS R5000 to MIPS R10000 implementation of MIPS architecture

| Feature | MIPS R5000 | MIPS R10000 | Diff. |
|-------------------|------------------|-------------------|-----------|
| Frequency | 180MHz | 200MHz | ~same |
| CMOS Process | 0.35 /3M | 0.35 /4M | |
| Cache Total | 32K/32K Cache | 32K/32KB Cache | ~same |
| Load/Store Unit | No | Yes | |
| Register Renaming | | Yes | |
| Peak Issue | 1+FP | 4 Issue | |
| Transistors | 3.6 Million | 5.9 Million | +64% |
| SPECint95 | 4.7 | 10.7 | +128% |
| SPECfp95 | 4.7 | 17.4 | +270% |
| Power | 10W | 30W | 200% |
| SPEC/Watt | 0.47/0.47 | 0.36/0.58 | -31%/ 23% |

Table 3. Comparison of Performance/Power and 1/Energy*Delay for representative RISC microporcessors

| Feature | Digital 21164 | MIPS 10000 | PowerPC 620 | HP 8000 | Sun UltraSpar |
|---|--------------------------|-------------------|------------------------|----------------|--------------------------|
| <i>Frequency</i> | 500 MHz | 200 MHz | 200 MHz | 180 MHz | 250 MHz |
| <i>Pipeline Stages</i> | 7 | 5-7 | 5 | 7-9 | 6-9 |
| <i>Issue Rate</i> | 4 | 4 | 4 | 4 | 4 |
| <i>Out-of-Order Exec.</i> | 6 loads | 32 | 16 | 56 | none |
| <i>Register Renam. (int/FP)</i> | none/8 | 32/32 | 8/8 | 56 | none |
| <i>Transistors/ Logic transistors</i> | 9.3M/ 1.8M | 5.9M/ 2.3M | 6.9M/ 2.2M | 3.9M*/ 3.9M | 3.8M/ 2.0M |
| <i>SPEC95 (Intg/FlPt)</i> | 12.6/18.3 | 8.9/17.2 | 9/9 | 10.8/18.3 | 8.5/15 |
| <i>Power</i> | 25W | 30W | 30W | 40W | 20W |
| <i>SpecInt/Watt</i> | 0.5 | 0.3 | 0.3 | 0.27 | 0.43 |
| <i>1/Energy*Delay</i> | 6.4 | 2.6 | 2.7 | 2.9 | 3.6 |

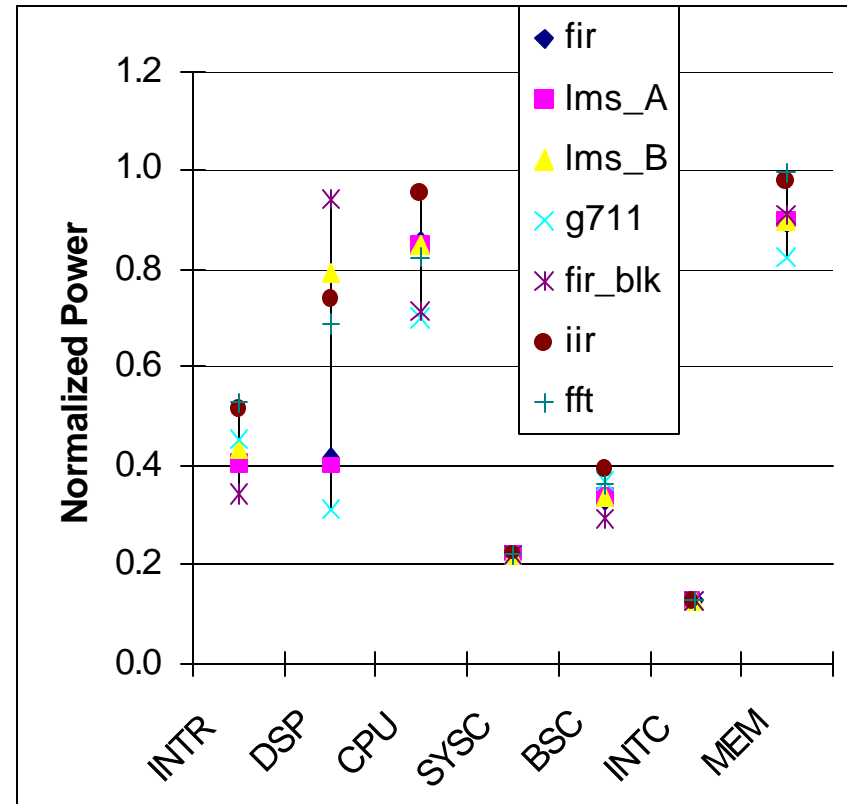
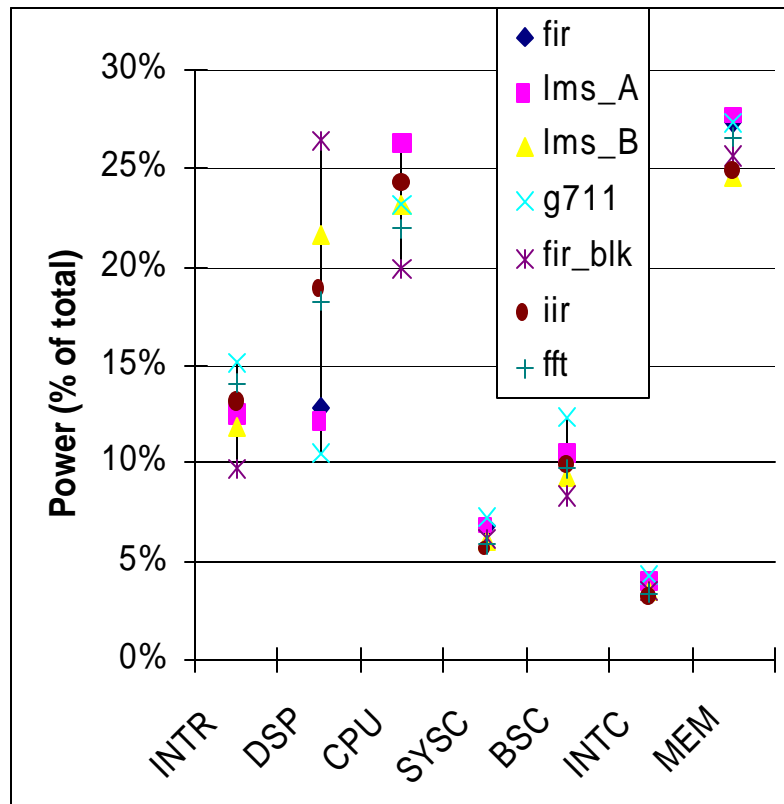
Table 4. A difference in power-performance factor resulting from doubling the size of caches

| Feature | 401 | 403 | Difference |
|--------------|--------------|------------------|------------------|
| Frequency | 50MHz | 66MHz (50MHz) | close |
| CMOS Process | 0.5u 3-metal | 0.5u 3-metal | same |
| Cache Total | 2K-I / 1K-D | 16K-I / 8K D | 8x |
| FPU | No | No | same |
| MMU | No | Yes | |
| Bus Width | 32 | 32 | same |
| Transistors | 0.3 Million | 1.82 Million | 600% |
| MIPS | 52 | 81 (61) | +56% (+17%) |
| Power | 140mW | 400mW (303mW) | +186% (+116%) |
| MIPS/Watt | 371 | 194 | -91% |

Table 5. The effect of doubling the caches in PowerPC architecture

| Feature | 604 | 620 | Difference |
|------------------------|----------------|-----------------|-------------------|
| <i>Frequency</i> | 100MHz | 133MHz (100MHz) | same |
| <i>CMOS Process</i> | 0.5u 4-metal | 0.5u 4-metal | same |
| <i>Cache Total</i> | 16K+16K Cache | 64K | ~double |
| <i>Load/Store Unit</i> | Yes | Yes | same |
| <i>Dual Intgr Unit</i> | Yes | Yes | same |
| <i>Reg- Renaming</i> | Yes | Yes | same |
| <i>Peak Issue</i> | 4 Instructions | 4 Instructions | same |
| <i>Transistors</i> | 3.6 Million | 6.9 Million | +92% |
| <i>SPECint92</i> | 160 | 225 (169) | +6% |
| <i>SPECfp02</i> | 165 | 300 (225) | +36% |
| <i>Power</i> | 13W | 30W (22.5W) | +73% |
| <i>Spec/Watt</i> | 12.3 / 12.7 | 7.5 / 10 | -64% |

Module-wise breakdown of the chip power consumption for the kernel benchmarks for the integrated RISC+DSP processor, (a) as a percentage of the total (b) normalized



Conclusions

- More specialized systems can benefit from re-configurable data-path designs. The main advantage is to reduce the clock and control overhead by mapping loops directly onto the re-configurable data-path.
- Where stream data or block data-processing dominates it makes sense to configure the data-path to compute algorithm specific operations.
- Aggressive use of chaining (as in vector processing) can be used to reduce memory accesses resulting in designs that may be called re-configurable vector pipelines.

Conclusions

- The best power-performance is obtained if the architecture is kept simple thus allowing improvements to be achieved by technology.
- We should seek improvements via simple design but increasing the clock frequency rather than keeping the frequency of operation low and increasing the complexity of the design.
- Power will be limiting performance of future processors.