Energy Considerations in Multichip-Module based Multiprocessors

James B. Burr and Allen M. Peterson

Space, Telecommunications, and Radioscience Laboratory
Department of Electrical Engineering
Stanford University
Stanford, Ca. 94305

Abstract
Multichip modules permit highly efficient implementation of “tiled” architectures. If the tiles are implemented in submicron CMOS, extremely high computation rates can be achieved, but power dissipation becomes the principal factor limiting achievable levels of integration and performance. This paper describes some examples of tiled architectures, and discusses the feasibility and advantages of reduced voltage operation for reducing energy per operation in power constrained environments.

1 Introduction
Signal processing algorithms are computationally intensive, and can often be implemented efficiently on massively parallel architectures. Energy per operation emerges as a dominant constraint in maximizing sustained performance. Multichip modules, tiled architectures, submicron CMOS, and low voltage operation are the four keys to maximizing performance in power-constrained applications.

2 Energy in digital systems
Electrical energy $E = \frac{1}{2}CV^2$, where $C$ is capacitance and $V$ is voltage. We have found it convenient to identify three distinct types of energy in digital systems. This classification arises from the distinctly different way each type of energy is optimized. The three types are:

- Decision energy: Energy consumed performing logical operations, such as control or datapath functions. Usually involves fairly localized communications: interconnect capacitance is significant, but gate capacitance is usually dominant.
- Storage energy: Energy required to store and retrieve data from memory elements. Often a considerable amount of power is dissipated in sense amplifiers in an effort to reduce access times. Leakage currents are very important since only a small percentage of transistors are switching at any given time.
- Communication energy: Energy required to transmit data from one place to another in a system. Principal components are wires, drivers, and receivers. Often a major source of power in a system, especially between packaged chips or between boards.

Minimizing power is not the same as minimizing energy per operation, since systems dissipate power through leakage currents even if they are doing no work. If power minimization is paramount, and there is not much computation or communication, then minimizing leakage is important. If, on the other hand, the system is computing a significant fraction of the time, then much larger leakage currents can be tolerated and may in fact lead to a reduction in overall power consumption.

3 Architecture, power, and technology scaling
The number of transistors which can be put on a single chip can easily exceed the amount of power available. For example (see Table 1), a 1 cm$^2$ chip can accommodate 400K transistors in 2µ CMOS. The
Table 2: Technology scaling rules; $S = \lambda_0 / \lambda$, the ratio of feature sizes in two technologies. Area power density is technology independent if the supply voltage scales as $1/S$.

<table>
<thead>
<tr>
<th>xstr/area</th>
<th>general</th>
<th>constV</th>
<th>scaleV</th>
</tr>
</thead>
<tbody>
<tr>
<td>cap/xstr</td>
<td>$S^2$</td>
<td>$S^2$</td>
<td>$S^2$</td>
</tr>
<tr>
<td>energy/xstr cap</td>
<td>$S^2V^2$</td>
<td>$S^2V$</td>
<td>$S^2V$</td>
</tr>
<tr>
<td>max freq</td>
<td>$S^3V^3$</td>
<td>$S^3V$</td>
<td>$S^3V$</td>
</tr>
<tr>
<td>ops/sec</td>
<td>$S/V^2$</td>
<td>$S$</td>
<td>$S^3$</td>
</tr>
<tr>
<td>power/xstr</td>
<td>$S^3V^3$</td>
<td>$S^3$</td>
<td>$S^3$</td>
</tr>
<tr>
<td>power/area</td>
<td>$S^3V^3$</td>
<td>$S^3$</td>
<td>$S^3$</td>
</tr>
</tbody>
</table>

Capacitive load of a 6:2 transistor is about 40fF; 10fF for the gate, 20fF for the source/drain, and 10fF for the interconnect. The number of transistors which can switch per second is $f_s = P_{ac}/(C_tV^2)$, where $P_{ac}$ is the average AC power consumption, $C_t$ is the average capacitance of a transistor, and $V$ is the voltage swing. If the power budget is 1W, and $V$ is 5V, then $f_s = 1/(25 \times 10^{-15}) = 10^{12}$ transistors per second. For $a = 10\%$, the maximum sustainable clock frequency is 25MHz.

In typical microprocessor applications, a large fraction of the transistors is dedicated to memory, so the overall activity ratio is less than 10%. To achieve 1 watt, 100MHz operation, a maximum of 100K transistors can be dedicated to computation on each cycle; the other 300K must be allocated to memory or idle resources.

Table 1 shows that even though constant voltage scaling (leaving voltage constant as technology scales) permits higher clock rates, it results in less performance at the same power level than if the voltage scales with the technology. This is because the total number of operations per second which can be supported at constant power is $f_{tech} \times xstr/\text{power}$, or $xstr/\text{energy}, S/V^3$. This means that, in computationally intensive power constrained applications, scaleV permits $S^3$ operations per second whereas constV permits only $S$.

In a power-constrained environment, architecture will be a function of technology if supply voltage does not scale. In constV, the maximum frequency sustainable switching all the transistors degrades substantially relative to the frequency sustainable in the technology. In scaleV, this ratio remains constant.

4 Multichip modules

Multichip modules offer the opportunity for substantially reduced communication energy by reducing system size, thereby decreasing capacitance. Volumetric processor density can be further increased by stacking modules as long as sufficient heat can be removed. Thermal management becomes much easier if power is reduced; fortunately power is a strong function of supply voltage.

4.1 Capacitance

Capacitance plays an important role in power dissipation, since logic decisions in CMOS involve moving charge from one place to another. Lowering the dielectric constant $\epsilon$ translates directly to power reduction since $P_{ac} = CV^2f$ and $C \propto \epsilon$. Reducing wire width, wire length, pad size, and via size can all help reduce power consumption.

A wire at any scale is about 1pF/cm. Principal techniques for reducing communication energy include reducing wire lengths, reducing voltage swings, and transmitting data only when it changes. Other forms of data compression are possible; the energy required to encode and decode a signal must be considered in evaluating overall energy efficiency.

4.2 Tiled architectures

We define a “tiled” architecture as consisting of a small number of VLSI die types designed to tessellate a plane. Multichip modules are an ideal vehicle for implementing tiled architectures because they permit closely spaced bare die and high I/O counts.

Systolic arrays [9, 10] and cellular automata [4, 18, 15] are two types of tiled architectures which have been investigated extensively in the literature. Systolic arrays are SIMD machines; cellular automata are MIMD machines. Most implementations of these architectures to date have consisted of large numbers of individually packaged chips placed on multiple circuit boards in racks of equipment with significant investments in power supplies and thermal engineering [5, 14, 3].

Several attempts have been made to build wafer-scale systems, with limited success due to yield problems. The opportunity exists to develop tile based architectures which map efficiently onto multichip modules. The size and composition of the tiles can be optimized to maximize overall system yield, and improve bare die testability.
Table 1: Computational capacity vs technology at 1 watt. \( f_{\text{max}} \) is the maximum frequency at which 10% of the devices could be switching. \( f_{\text{tech}} \) is the maximum frequency assuming a logic depth of 10. \( \maxc \) is the maximum number of transistors which can be devoted to concurrent computation at \( f_{\text{tech}} \). \( c \) is \( \maxc \) as a percentage of the total number of transistors. Constant voltage scaling requires a decreasing percentage of system resources to be active, implying an architecture change. Scaling the supply voltage permits the architecture to scale as well.

\[
\begin{array}{|c|c|c|c|c|c|c|c|c|c|}
\hline
\text{tech} & \text{xstr/cm}^2 & \text{vdd} & f_{\text{F/xstr}} & f_{\text{J/xstr}} & f_{\text{xstr/sec}} & f_{\text{max}} & f_{\text{tech}} & \maxc & c \\
\hline
\text{constV} & 2.0 & 400K & 5.00 & 40 & 1000 & 1e12 & 25.00 & 100 & 100K & 25 \\
1.0 & 1.6M & 5.00 & 20 & 500 & 2e12 & 12.50 & 250 & 80K & 5 \\
0.5 & 6.4M & 5.00 & 10 & 250 & 4e12 & 6.25 & 800 & 50K & 0.7 \\
\hline
\text{scaleV} & 2.0 & 400K & 5.00 & 10 & 250 & 4e12 & 6.25 & 800 & 50K & 0.7 \\
1.0 & 1.6M & 2.50 & 10 & 125 & 8e12 & 50.00 & 200 & 400K & 25 \\
0.5 & 6.4M & 1.25 & 10 & 16 & 64e12 & 100.00 & 400 & 1.6M & 25 \\
\hline
\end{array}
\]

A number of neural network chips have been reported which are well suited for tiling on multichip modules. The Adaptive Solutions CNAPS [6], Neural Semiconductor’s DNNA [8], and Bellcore’s CLC [1], are all cascadable architectures.

Each CNAPS chip dissipates 4 watts and performs 1 GCPS (billion connections per second); a 64 chip array on a 10x10 cm multichip module would dissipate 256 watts and perform 64 GCPS. If Vdd were reduced to 1V, the same module would achieve 12.8 GCPS and dissipate only 2 watts.

In a large scale programmable logic application, the Quickturn PCB contains 63 Xilinx chips implemented on a 18.5" x 19" circuit board with 14 signal layers. Most of the power dissipation in each Xilinx chip comes from driving chip I/Os. Assuming 50pF per pad, 50 outputs, and 10% switching at 50MHz, \( aNCV^2f = 0.1 \times 50 \times 50e-12 \times 25 \times 50e6 = 300mW \) per chip, or 20 watts overall. If implemented on a 9x7cm MCM substrate, and if the pads were optimized to drive multichip module loads, the load capacitance could drop by a factor of 5, reducing the power to 4 watts overall. If the supply voltage were reduced to 1 volt, performance would degrade by a factor of 5 but the entire module would dissipate only 32mW.

Submicron CMOS offers increased density and reduced gate delay. Although the energy per operation is reduced, the area power density increases unless the supply voltage is also reduced. Submicron RISC microprocessor implementations have taken advantage of the additional transistors by increasing cache size and integrating floating point coprocessors on the same chip. Real estate is devoted to resources which are inactive most of the time. This reduces the system-wide activity ratio \( a \), partially offsetting the increased power density inherent in the technology.

Scalable architectures, on the other hand, feel the full effects of the increased power density since \( a \) remains constant as more resources are added as the technology scales down. If a scalable architecture is designed to tile the surface of a multichip module, power management becomes a major problem, and a reduction in supply voltage becomes mandatory.

5 Low voltage digital logic

Low supply voltages normally require reduced device thresholds. The lower bound on threshold voltage is determined by the leakage current of OFF devices. In standard CMOS processes, the leakage current is about 1pA (10\(^{-12}\) amps), so the DC power dissipation is about 5pW per device at 5 volts. One of the major attractions of CMOS has been its extremely low DC power dissipation.

Computational circuits which are in use most of the time do not need such low DC power dissipation. These circuits can be built using devices which have much lower thresholds. A reasonable design criterion is to constrain the DC power dissipation to be less than the AC power dissipation. This requirement can be expressed in terms of the ratio of the ON current \( I_{\text{on}} \) to the OFF current \( I_{\text{off}} \) of a device. If \( Id \) is the logic depth, the number of stages of logic between pipe stages, and \( a \) is the activity ratio, the fraction of gates which are switching at any given time, then since \( P_{\text{ac}} = aCV^2f \), \( P_{\text{dc}} = I_{\text{off}}V \), and \( I_{\text{on}} = ldCVf \), \( I_{\text{on}}/I_{\text{off}} = ld/a \). If pipe stages are placed every 10 stages of logic, and if 10% of the gates are switch-
ing at a given time, then \( Id = 10 \) and \( a = 0.1 \), so \( I_{on}/I_{off} = 100 \).

The subthreshold current of an MOS transistor \( (I_{ds} \text{ when } V_{gs} < V_{t}) \) is \([16, 13, 17, 11]\)

\[
I_{ds} = knV_{t}^{2}e^{(V_{gs}-V_{t})/(nV_{T})}(1 - e^{-V_{ds}/V_{T}})
\]

where

\[
k = u_{n}C_{ox}W/L
\]

\[
n = 1 + \Omega T_{ox}/d_{0}
\]

\[
V_{T} = KT/q
\]

\[
C_{ox} = \epsilon_{ox}/T_{ox}
\]

\[
\Omega = \epsilon_{si}/\epsilon_{ox}
\]

and \( I_{ds} \) is drain current, \( A \); \( V_{gs} \) is gate-source voltage, \( V \); \( V_{T} \) is thermal voltage, \( V \); \( V_{t} \) is threshold voltage, \( V \); \( \epsilon \) is carrier mobility, \( \text{cm}^{2}/\text{Vsec} \); \( W \) is transistor width, \( \text{cm} \); \( L \) is transistor length, \( \text{cm} \); \( d_{0} \) is channel depletion width, \( \text{cm} \); \( T_{ox} \) is gate oxide thickness, \( \text{cm} \); \( C_{ox} \) is gate capacitance, \( \text{F/cm}^{2} \); \( K \) is Boltzmann’s constant, \( \text{V/degK} \); \( T \) is temperature, \( \text{degK} \); and \( q \) is electronic charge, coulombs.

\( V_{T} \) is 0.026mV at \( T = 300 \) deg K, so for \( V_{ds} > V_{T} \), \( e^{-V_{ds}/V_{T}} \approx 0 \), and

\[
I_{ds} = knV_{t}^{2}e^{(V_{gs}-V_{t})/(nV_{T})}
\]

This expression is log linear. To find the slope, find the ratio of two currents:

\[
I_{1}/I_{2} = e^{(V_{1}-V_{2})/(nV_{T})}
\]

so

\[
m_{s} = \frac{\Delta V}{\Delta \log_{10}(I)} = nV_{T} \ln(10)
\]

This is the subthreshold slope, expressed in mV/decade of current. In \( 2\mu \) CMOS, \( n \) is about 1.4, so \( m_{s} \) is about 80mV/decade. At \( V_{gs} = V_{t} \),

\[
I_{0} = knV_{t}^{2}
\]

In \( 2\mu \) CMOS, \( u_{n} = 600 \) \text{cm}^{2}/\text{Vsec} \), and \( T_{ox} = 40 \text{nm} \), so if \( W = 4\mu \) and \( L = 2\mu \), then \( k = 103.5\mu \text{A/V}^{2} \) and \( I_{0} = 0.1\mu \text{A} \).

A logic depth of 10 and an activity ratio of 0.1 support a current ratio of 100, requiring a \( \Delta V \) of 160mV at 80mV/decade. If \( V_{t} \) were set to 160mV, a single transistor would sink 1nA at 0V.

Suppose we want to sink 1pA at \( V_{dd} \) with \( V_{t} = 160 \)mV. For a transistor in saturation \( (V_{gs} < V_{ds} + V_{t}) \),

\[
I_{ds} = \frac{k}{2}(V_{gs} - V_{t})^{2}
\]

To find the supply voltage which can provide the desired current,

\[
V_{dd} = V_{t} + \sqrt{2I_{ds}/k}
\]

For \( I_{ds} = 1\mu \text{A} \), \( V_{dd} = 300 \)mV in \( 2\mu \) CMOS.

If \( V_{t} \) can be adjusted to 160mV, and if \( V_{dd} \) is set to 300mV, transistors will sink 1nA when they are off and 1\mu A when they are on. The DC power dissipation will be 300pW \((1\mu \text{A} \times 300 \text{mV}) \). The AC power dissipation is given by \( aC_{g}V^{2}f \), and \( f = I_{on}/(IdC_{g}V) \), where \( C_{g} \) is the capacitance of a single gate. Usually \( C_{g} = 4C_{t} \), so \( f = I_{on}/(4IdC_{t}V) \). Then

\[
P_{ac} = \frac{(a/Id)I_{on}V}{(a/Id)P_{on}}
\]

where \( P_{on} \) is the power dissipation of a transistor when it is fully on.

In \( 2\mu \) CMOS, if \( V_{dd} = 300 \)mV, \( a = 0.1 \), and \( ld = 10 \), then \( f = 2 \text{MHz} \), and \( P_{ac} = 3 \) nW per transistor. The energy per transistor is \( \frac{1}{2}C_{t}V^{2} = 1.8 \text{fJ} \).

From above, a 5-bit multiplier consumes 60\text{N}^{2}\text{V}^{2}/\text{S/fJ}. At 300mV, in \( 2\mu \) CMOS, this multiplier would consume 135 fJ. In 0.25\mu CMOS, the same multiplier would require 17fJ per connection. Operating frequency scales as \( S^{2}/V \), so it would run at 2 MHz \( \times(2/0.25)^{2} \) or 128MHz.

Classical technology scaling becomes less valid as feature sizes shrink due to fixed potentials built into the materials, especially short channel effects. It applies more completely to low-V\(_{t}\) processes because the onset of short-channel effects is delayed by the reduced electric field strength.

In a low-V\(_{t}\) process, \( I_{eff} \) cannot be neglected in computing total energy.

\[
t_{pd} = Q/I = CV/I
\]

\[
E_{dc} = I_{eff}V_{pd}
\]

\[
= CV^{2}/(I_{on}/I_{off} - 1)
\]

\[
E_{ac} = \frac{1}{2}CV^{2}
\]

\[
E = E_{ac} + E_{dc}
\]

This implies \( E \) is a minimum when \( I_{on}/I_{off} \) is a maximum. \( I_{on}/I_{off} \) is maximum in the subthreshold region since \( I_{ds} \) is an exponential function of voltage below \( V_{t} \) and a quadratic function of voltage above \( V_{t} \). Since subthreshold current is log linear, energy is constant throughout the subthreshold region. That is, if \( V_{dd} \leq V_{t} \), the optimum placement of \( V_{gs} \) is at
Energy vs Vdd in the subthreshold region. DC energy decreases exponentially as Vdd increases. AC energy increases quadratically as Vdd increases. The minimum is around Vdd = 200mV.

\[ V_t - V_{dd}, \text{ since energy is constant for any } V_{ss} < V_t - V_{dd} \text{ and performance is maximized when } V_{dd} = V_t. \]

In 2.0\( \mu \) CMOS, \( I_0 = 0.1\mu A \), so if \( V_{dd} = V_t = 160mV \), then \( f = 390KHz, P_{ac} = 160pW, \) and \( E_{ac} = 0.5fJ \). A 5-bit multiplier would consume 38fJ. In 0.25\( \mu \) CMOS, the same multiplier would require 4.8fJ, and would run at 390KHz \times (2/25)^2, or 25MHz.

5.1 Previous work

Low voltage digital logic is not new. In 1974, Richard Swanson built an 11-stage ring oscillator and ran it at voltages down to 100mV [16]. In 1985, Eric Vittoz published a paper describing a variety of micropower circuit techniques, including micropower digital logic [17]. Vittoz was reporting techniques which are widely used in making digital watches, where power dissipation is an overriding constraint. Our emphasis is somewhat different, in that we seek to minimize the energy per operation in order to maximize performance on a finite, if sizeable, power budget. 3D stacking of multichip modules will eventually permit 800cm\(^2\) of silicon per cubic inch. If each cm\(^2\) contains a RISC processor that dissipates one watt, it will be necessary to remove around 1 kW/in\(^3\) from the system. One popular alternative is to engineer systems which can cope with that kind of power density. Another alternative worth considering is to reduce the supply voltage to reduce power. Because performance in proportional to \( V \) and power to \( V^3 \), even a small reduction in supply voltage can lead to large power savings.

5.2 Modeling

Existing current models have several characteristics which make modeling near \( V = V_t \) difficult. In particular, the standard form saturation equation goes to zero at \( V = V_t \) (see Figure 2) so there is a discontinuity for 100mV or so above \( V_t \) which in real transistors is a gradual transition region between weak and strong inversion, diffusion and drift current.

We have been using MATLAB to compute the current-voltage characteristics of various circuits, including an inverter, a nand gate, and several different flavors of latches. This gives us low voltage decision and storage elements from which we can construct digital systems. We wired up an inverter using discrete transistors and measured its voltage transfer characteristic to verify our results, obtaining good agreement with our MATLAB models.

5.3 Low voltage process opportunities

Standard CMOS processes are not optimized for low-voltage operation. To prevent punchthrough at short channel lengths and high source/drain electric fields, channel doping must be increased, which increases the subthreshold slope (in mV/decade). While
a lower bound of 80mV/decade is achievable at room temperature \(dV = nT_\text{c} \ln(10)\) with \(n = 1\). \(dV\) is more typically 80mV/decade in 2\(\mu\) CMOS and 90mV/decade in 0.8\(\mu\) CMOS. \(T_\text{c}/dV\) can be reduced by reducing \(N_B\), since \(dV = \sqrt{2\phi} \ln(N_B/n_i)\), where \(\phi = V_T \ln(N_B/n_i)\) and \(n_i = \sqrt{1.57}\times e^{-1.15/V_F} \times 10^{16}\) [13].

Low gate, drain, and threshold voltages permit all doping concentrations to be reduced, since electric field strength at 200mV is 1/25 the field strength at 5V. This has two benefits for low voltage operation:

First, \(n\) is reduced, decreasing the subthreshold slope and thus reducing the supply voltage (and therefore energy per operation) necessary to achieve the desired on/off current ratio.

Second, source/drain capacitances are reduced, further reducing energy per operation.

### 5.4 Barriers to practical implementation

Several problems have hampered widespread use of micropower CMOS. These are 1) external interfacing, 2) controlling \(V_t\), 3) noise margins, 4) power consumption of OFF devices, 5) slow circuits, and 6) nonuniform activity ratios.

#### 5.4.1 External interfacing

Pressures to maintain compatibility with the external world will require energy efficient level-shifting interface circuits. Multichip modules make it easier to isolate a low-voltage subsystem from the external world. Interface circuits can be placed on the periphery of the module and electrically isolated from the rest of the system just as digital circuitry can be isolated from low-noise analog subsystems.

Some MCM technologies provide intrinsic bypass capacitance by placing a thin (138nm in the ATT process) dielectric between Vdd and Gnd planes. This, together with area bonding, can deliver charge locally throughout the chip to satisfy peak current demands. Charge stored is \(CV\); current is \(\frac{1}{2}(V - V_f)^2\), so the ability of the system to meet peak current demand improves at reduced voltage.

#### 5.4.2 Controlling \(V_t\)

Just how far the supply voltage can be lowered depends on controllability of \(V_t\). Swanson asserts \(V_t\) can be controlled within 2% [16]. Then

\[
\frac{I_+}{I_-} = e^{\frac{\phi}{V_F}}
\]

\[
e^{0.02 \times 200 / 0.20 / 1.4}
\]

More recently, Pavasovic et al. have characterized CMOS process variations by measuring variations in subthreshold current, and have observed spatially periodic fluctuations in subthreshold current of up to 30% of the average current [12].

In standard processes, poly is usually doped n+. This tends to reduce \(V_{th}\) and increase \(V_t\). Without a channel implant, \(V_{th} \approx 0\) V, and \(V_t \approx -1.4\) V. P-type channel implant is usually used to reduce \(|V_{th}|\), further reducing hole mobility.

The Stanford BiCMOS process dope n+ gates n+ and p+ gates p+, resulting in low intrinsic thresholds for both types of devices. With low intrinsic thresholds, the substrate bias voltage \(V_s\) can be biased to adjust \(V_t\), since \(V_t = V_{th} + \gamma \sqrt{2\phi + V_s} - \frac{1}{2} |\phi_F|\) [7].

#### 5.4.3 Noise margins

Internal noise sources scale at least as fast as the supply voltage: capacitive coupling scales as \(V^2\); inductive coupling as \(V^3\); resistive coupling as \(V^2\). Thermal noise is about 100 microvolts \((V^2 = 4KTR_f)\) [2].

Although individual device noise will be small, current transients could be large as in CMOS; this can be mitigated by embedding bypass capacitors on chip in the Vdd/Gnd rails, using solder-bumped Vdd and Gnd pads throughout the chip, and an MCM technology like ATT’s with 25nF/cm intrinsic bypass capacitance between the Vdd and Gnd planes.

External noise is less well characterized. Low voltage circuits will have to be well insulated from parts of the system operating at higher voltages/currents. Power supply noise will be a major problem. One watt at 200mV will require 5 amps.

#### 5.4.4 Power consumption of OFF devices

Assuming 100mV/decade in the region around \(V_t\), and 100nA at \(V_t\), an OFF current of 1nA implies a \(V_t\) of 200mV. An ON current of 1\(\mu\)A implies a supply voltage of 300mV (100mV above \(V_t\)).

1nA per gate at 200mV implies 200\(\mu\)W per gate DC. One million gates would dissipate 200\(\mu\)W. If the same one million gates are switching at 10MHz, and assuming 40\(\Omega\) per gate in 0.5\(\mu\) CMOS (10\(\Omega\)/transistor), and assuming 10% of the gates switch, we get

\[
N_g IV = 1e6 \times 1e-9 \times 0.2 = 200\mu W \text{ DC}
\]
so the DC power would be 200µW and the AC power 1.6mW. Assuming 100 gates per operation (as in 4 bit precision neural net computation) would give 1e11 ops/sec at 1.8mW or 18fJ/op.

This approach would not be as good if most of the gates were idle, but has great promise for massively parallel architectures. A special problem is that dynamic memory should still be constructed with high threshold devices to reduce leakage currents and dissipate less DC power. Low power on-chip interface circuits will be required between computational and storage elements. Leakage current could be further reduced by disconnecting the power supply from idle resources.

5.4.5 Slow circuits

Although the circuits are slow, they are not as slow as they are low power. Speed scales as V, energy as V^2, and power as V^3, so a modest reduction in speed results in a large reduction in both energy and power. Aggressive voltage reduction is not interesting for uniprocessors, which only need voltage to scale with S to maintain constant power per unit area; but is very interesting for massively parallel architectures. The optimum supply voltage will be determined largely by the power budget and the cost of heat removal.

5.4.6 Nonuniform activity ratios

The optimum supply voltage depends on the activity ratio a. However, because OFF current is exponential with voltage, only small changes in supply voltage are necessary to accommodate large variations in activity ratios. The only reason to turn devices off hard is to permit dynamic circuits to run at low frequency. In particular, dynamic RAMs will need much higher refresh rates. A 100fF storage node will leak away in 20µsec if OFF current is 1nA. The optimum supply voltage for a DRAM will occur approximately where the refresh power equals the read/write power, since

\[ P_{\text{refresh}} \propto V e^{-V} \text{ and } P_{\text{rw}} \propto V^3. \]

Level shifting interface circuits will be unavoidable in systems combining high computation rates and large amounts of memory, since decision and storage elements have different optimal supply voltages.

6 Applications: Petacube

One possible scenario could be a MOSIS Tincchip (2mm x 2mm as the basic building block, (or whatever size optimizes overall system yield) tiled over a set of stacked MCMs. One cubic inch could support 800cm^2 of tinytiles; at 1 million gates/cm^2 in 0.5µ CMOS, 100 gates per operation, and 200mV, the cube would dissipate 1.4W and perform 1e14 operations per second. At 1 volt the same cube would dissipate 175 watts, and perform 5e14 operations per second. At 3.3 volts the same cube would dissipate 4.5KW, and perform 1.3e15 operations per second. Figure 3 shows power vs supply voltage.

7 Conclusion

The combination of tiled architectures, submicron CMOS, multichip modules, and low voltage operation can greatly enhance system performance in power constrained applications. Although reducing the supply voltage lowers performance by V, it reduces energy by V^2 and power by V^3. This becomes especially important as technology scales down, since performance increases as S^2V. Some important open issues are

- Practical limits to low voltage operation require further investigation.
- Tile size, composition, and variety should be chosen to maximize overall system yield. This will likely result in novel partitioning schemes, and may even reverse the trend toward ever-increasing die sizes.
• 100% built-in self test is an important goal, and should be an architecture driver in system partitioning.

• 3D stacked architectures will have far fewer thermal problems at low voltage.

Energy per operation is the single most important criterion in designing high performance, massively parallel machines on limited power budgets. Multichip modules can play a significant role in system architectures by reducing communication energy in digital systems. Furthermore, they encourage a reevaluation of traditional partitioning.

Acknowledgements

This research was supported in part by NASA grants NAGW1910 and NAGW419, by a gift from Intel Corporation, and by a grant from Stanford’s Center for Integrated Systems. Multichip modules were provided by ATT, workstations by Sun Microsystems, and VLSI fabrication by MOSIS.

References


