Panel Session:  
*Amending Moore’s Law for Embedded Applications*

James C. Anderson  
MIT Lincoln Laboratory

HPEC04  
29 September 2004

This work is sponsored by the HPEC-SI (high performance embedded computing software initiative) under Air Force Contract F19628-00-C-0002. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the United States Government.

Reference to any specific commercial product, trade name, trademark or manufacturer does not constitute or imply endorsement.
Objective, Questions for the Panel & Schedule

- **Objective**: identify & characterize factors that affect the impact of Moore’s Law on embedded applications

- **Questions for the panel**
  - 1). Moore’s Law: what’s causing the slowdown?
  - 2). What is the contribution of Moore’s Law to improvements at the embedded system level?
  - 3). Can we preserve historical improvement rates for embedded applications?

Panel members & audience may hold diverse, evolving opinions

- **Schedule**
  - 1540-1600: panel introduction & overview
  - 1600-1620: guest speaker Dr. Robert Schaller
  - 1620-1650: panelist presentations
  - 1650-1720: open forum
  - 1720-1730: conclusions & the way ahead
Panel Session: 
*Amending Moore’s Law for Embedded Applications*

**Moderator:** Dr. James C. Anderson, MIT Lincoln Laboratory

**Dr. Richard Linderman,**
Air Force Research Laboratory

**Dr. Mark Richards,**
Georgia Institute of Technology

**Mr. David Martinez,**
MIT Lincoln Laboratory

**Dr. Robert R. Schaller,**
College of Southern Maryland
Four Decades of Progress at the System Level

Gordon Moore publishes “Cramming more components onto integrated circuits”

Computers lose badly at chess
Four Decades of Progress at the System Level

1965

Gordon Moore publishes “Cramming more components onto integrated circuits”

1997

Computers lose badly at chess

Robert Schaller publishes “Moore’s Law: past, present and future”

Deep Blue (1270kg) beats chess champ Kasparov
Four Decades of Progress at the System Level

1965

Gordon Moore publishes “Cramming more components onto integrated circuits”

1997

Robert Schaller publishes “Moore’s Law: past, present and future"

Deep Blue (1270kg) beats chess champ Kasparov

2002-2004

Mark Richards (with Gary Shaw) publishes “Sustaining the exponential growth of embedded digital signal processing capability”

Chess champ Kramnik ties Deep Fritz & Kasparov ties Deep Junior (10K lines C++ running on 15 GIPS server using 3 Gbytes)
Four Decades of Progress at the System Level

1965
Gordon Moore publishes “Cramming more components onto integrated circuits”

1997
Robert Schaller publishes “Moore’s Law: past, present and future”

2002-2004
Mark Richards (with Gary Shaw) publishes “Sustaining the exponential growth of embedded digital signal processing capability”

~2005
Deep Dew handheld chess champ (0.6L & 0.6kg) uses 22 AA cells (Li/FeS₂, 22W for 3.5 hrs) & COTS parts incl. voice I/O chip

Computers lose badly at chess

Deep Blue (1270kg) beats chess champ Kasparov

Chess champ Kramnik ties Deep Fritz & Kasparov ties Deep Junior (10K lines C++ running on 15 GIPS server using 3 Gbytes)
Four Decades of Progress at the System Level

1965
Gordon Moore publishes "Cramming more components onto integrated circuits"

1997
Robert Schaller publishes "Moore's Law: past, present and future"

Deep Blue (1270kg) beats chess champ Kasparov

2002-2004
Mark Richards (with Gary Shaw) publishes "Sustaining the exponential growth of embedded digital signal processing capability"

Chess champ Kramnik ties Deep Fritz & Kasparov ties Deep Junior (10K lines C++ running on 15 GIPS server using 3 Gbytes)

~2005
Deep Dew hand-held chess champ (0.6L & 0.6kg) uses 22 AA cells (Li/FeS2, 22W for 3.5 hrs) & COTS parts incl. voice I/O chip

~2008
Deep Yogurt has 1/3 the size & power of Deep Dew, with 3X improvement in 3 yrs
Power per Unit Volume (Watts/Liter) for Representative Systems ca. 2003

Throughput in GIPS (billions of Dhrystone instructions/sec)

Deep Fritz & Deep Junior Chess Server

70 W/L limit for convection-cooled cards (typical for conduction-cooled)

Hand-held unit feasible with COTS parts 4Q03, but not built

70 W/L limit for convection-cooled cards (typical for conduction-cooled)
Power per Unit Volume (Watts/Liter) for Representative Systems ca. 2003

Throughput in GIPS (billions of Dhrystone instructions/sec)

Computation Density (GIPS/Liter)

Deep Fritz & Deep Junior Chess Server

Chess champs brains

Human chess champs Kramnik & Kasparov

Kramnik & Deep Fritz

Hand-held unit feasible with COTS parts 4Q03, but not built

1.6 W/L, moderately active human (human vs. machine "Turing Tests")

70 W/L, limit for convection-cooled cards (typical for conduction-cooled)
Power per Unit Volume (Watts/Liter) for Representative Systems ca. 2003

Throughput in GIPS (billions of Dhrystone instructions/sec)

- PowerPC 750FX (0.13μm, 800 MHz)
- Active die volume (1μm depth)
- Die (1mm thick)
- Packaged device

- Computation Density (GIPS/Liter)
- Computation Efficiency (GIPS/watt)

- RBMK-1500 reactor
- 8000 Watts/Liter
- 70 W/L limit for convection-cooled cards (typical for conduction-cooled)
- 1.6 W/L, moderately active human
- Hand-held unit feasible with COTS parts 4Q03, but not built

- Human chess champs
- Kramnik & Kasparov

- Deep Fritz & Deep Junior Chess Server

- Computer card

- 70 W/L, moderately active human
- Human vs. machine “Turing Tests”

- Kramnik & Deep Fritz
System-level Improvements Falling Short of Historical Moore’s Law

GFLOPS (billions of 32 bit floating-point operations/sec) sustained for 1K complex FFT using 6U form factor convection-cooled COTS multiprocessor cards <55W, 2Q04 data.

GFLOPS (billions of 32 bit floating-point operations/sec) sustained for 1K complex FFT using 6U form factor convection-cooled COTS multiprocessor cards <55W, 2Q04 data.

COTS ASIC & FPGA improvements outpacing general-purpose processors, but all fall short of historical Moore’s Law.

Moore’s Law slope: 100X in 10 yrs.
Pair of analog-to-digital converters provide data to processor card for 32 bit floating-point 1K complex FFT

Highest-performance 6U form factor multiprocessor cards <55W

Moore’s Law slope: 4X in 3 yrs
General-purpose µP, DSP & RISC (w/ vector processor)

12- to 14-bit ADCs

1 X in 3 yrs
Special-purpose ASICs
SRAM-based FPGAs

3 X in 3 yrs
SRAM-based FPGAs

2 X in 3 yrs
SRAM-based FPGAs

Open systems architecture goal: mix old & new general- & special-purpose cards, with upgrades as needed (from 1992-2003, a new card could replace four 3-yr-old cards)

Projections assume future commercial market for 1 GSPS 12-bit ADCs & 50 GFLOPS cards with 8 Gbytes/sec I&O
Representative Embedded Computing Applications

Sonar for anti-submarine rocket-launched lightweight torpedo (high throughput requirements but low data rates)

Radio for soldier’s software-defined comm/nav system (severe size, weight & power constraints)

Radar for mini-UAV surveillance applications (stressing I/O data rates)

~3m wingspan

Cost- & schedule-sensitive real-time applications with high RAS (reliability, availability & serviceability) requirements
Embedded Signal Processor Speed & Numeric Representations Must Track ADC Improvements

Sonar example near limit of 32 bit floating-point (18 ADC bits @ 100 KSPS + 5 bits processing gain vs. 23 bit mantissa + sign bit)

Radio example near limit of 16 bit fixed-point (10 ADC bits @ 400 MSPS + 5 bits processing gain)

*Floating-point preferred (same memory & I/O as fixed-point)
Objective, Questions for the Panel & Schedule

• Objective: identify & characterize factors that affect the impact of Moore’s Law on embedded applications

• Questions for the panel
  – 1). Moore’s Law: what’s causing the slowdown?
  – 2). What is the contribution of Moore’s Law to improvements at the embedded system level?
  – 3). Can we preserve historical improvement rates for embedded applications?

Panel members & audience may hold diverse, evolving opinions

• Schedule
  – 1540-1600: panel introduction & overview
  – 1600-1620: guest speaker Dr. Robert Schaller
  – 1620-1650: panelist presentations
  – 1650-1720: open forum
  – 1720-1730: conclusions & the way ahead
Objective, Questions for the Panel & Schedule

- **Objective:** identify & characterize factors that affect the impact of Moore’s Law on embedded applications

- **Questions for the panel**
  - 1). Moore’s Law: what’s causing the slowdown?
  - 2). What is the contribution of Moore’s Law to improvements at the embedded system level?
  - 3). Can we preserve historical improvement rates for embedded applications?

**Panel members & audience may hold diverse, evolving opinions**

- **Schedule**
  - 1540-1600: panel introduction & overview
  - 1600-1620: guest speaker Dr. Robert Schaller
  - 1620-1650: panelist presentations
  - 1650-1720: open forum
  - 1720-1730: conclusions & the way ahead
Conclusions & The Way Ahead

- Slowdown in Moore’s Law due to a variety of factors
  - Improvement rate was 4X in 3 yrs, now 2-3X in 3 yrs (still substantial)
  - Impact of slowdown greatest in “leading edge” embedded applications
  - Software issues may overshadow Moore’s Law slowdown
- COTS markets may not emerge in time to support historical levels of improvement
  - Federal government support may be required in certain areas (e.g., ADCs)
  - Possible return of emphasis on advanced packaging and custom devices/technologies for military embedded applications
- Developers need to overcome issues with I/O standards & provide customers with cost-effective solutions in a timely manner: success may depend more on economic & political rather than technical considerations
- Hardware can be designed to drive down software cost/schedule, but new methodologies face barriers to acceptance
- Improvements clearly come both from Moore’s Law & algorithms, but better metrics needed to measure relative contributions

“It’s absolutely critical for the federal government to fund basic research. Moore’s Law will take care of itself. But what happens after that is what I’m worried about.”
- Gordon Moore, Nov. 2001
Points of Reference

• **6U form factor card**
  - Historical data available for many systems
  - Convection cooled
    - Fans blow air across heat sinks
    - Rugged version uses conduction cooling
  - Size: 16x23cm, 2cm slot-to-slot (0.76L)
  - Weight: 0.6kg, typ.
  - Power: 54W max. (71W/L)
    - Power limitations on connectors & backplane
    - Reliability decreases with increasing temperature
  - Can re-package with batteries for hand-held applications (e.g., walkie-talkie similar to 1L water bottle weighing 1kg)

• **1024-point complex FFT (fast Fourier transform)**
  - Historical data available for many computers (e.g., fftw.org)
  - Realistic benchmark that exercises connections between processor, memory and system I/O
  - Up to 5 bits processing gain for extracting signals from noise
  - Expect 1µsec/FFT (32 bit floating-point) on 6U COTS card ~7/05
    - Assume each FFT computation requires 51,200 real operations
    - 51.2 GFLOPS (billions of floating point operations/sec) throughput
    - 1024 MSPS (million samples samples/sec, complex) sustained, simultaneous input & output (8 Gbytes/sec I&O)
Moore’s Law & Variations, 1965-1997

• “Original” Moore’s Law (1965, revised 1975)
  – 4X transistors/die every 3 yrs
  – Held from late ’70s - late ’90s for DRAM (dynamic random access memory), the most common form of memory used in personal computers
  – Improvements from decreasing geometry, “circuit cleverness,” & increasing die size
  – Rates of speed increase & power consumption decrease not quantified

• “Amended” Moore’s Law: 1997 National Technology Roadmap for Semiconductors (NTRS97)
  – Models provided projections for 1997-2012
  – Improvement rates of 1.4X speed @ constant power & 2.8X density (transistors per unit area) every 3 yrs
  – For constant power, speed x density gave max 4X performance improvement every 3 yrs
  – Incorrectly predicted 560 mm² DRAM die size for 2003 (4X actual)

Historically, Performance = $2^{\frac{\text{Years}}{1.5}}$
Moore’s Law Slowdown, 1999-2003
(recent experience with synchronous DRAM)

- **Availability issues**: production did not come until 4 yrs after development for 1Gbit DDR (double data rate) SDRAMs (7/99 – 7/03)
- **SDRAM price crash**
  - 73X reduction in 2.7 yrs (11/99 – 6/02)
  - Justice Dept. price-fixing investigation began in 2002
- **Reduced demand**
  - Users unable to take advantage of improvements as $3 SDRAM chip holds 1M lines of code having $100M development cost (6/02)
  - Software issues made Moore’s Law seem irrelevant
    - Moore’s Law impacted HW, not SW
    - Old SW development methods unable to keep pace with HW improvements
    - SW slowed at a rate faster than HW accelerated
    - Fewer projects had HW on critical path
    - In 2000, 25% of U.S. commercial SW projects ($67B) canceled outright with no final product
    - 4 yr NASA SW project canceled (9/02) after 6 yrs (& $273M) for being 5 yrs behind schedule

System-level improvement rates possibly slowed by factors not considered in Moore’s Law “roadmap” models
The End of Moore’s Law, 2004-20XX

• 2003 International Technology Roadmap for Semiconductors (ITRS03)
  – Models provide projections for 2003-2018
  – 2003 DRAM size listed as 139 mm² (1/4 the area predicted by NTRS97)
  – Predicts that future DRAM die will be smaller than in 2003
  – Improvement rates of 1.5X speed @ constant power & 2X density every 3 yrs
  – Speed x density gives max 3X performance improvement every 3 yrs
  – Limited by lithography improvement rate (partially driven by economics)

• Future implications (DRAMs & other devices)
  – Diminished “circuit cleverness” for mature designs (chip & card level)
  – Die sizes have stopped increasing (and in some cases are decreasing)
  – Geometry & power still decreasing, but at a reduced rate
  – Fundamental limits (e.g., speed of light) may be many (more) years away
    Nearest-neighbor architectures
    3D structures
  – Heat dissipation issues becoming more expensive to address
  – More chip reliability & testability issues
  – Influence of foundry costs on architectures may lead to fewer device types
    in latest technology (e.g., only SDRAMs and static RAM-based FPGAs)

Slower (but still substantial) improvement rate predicted, with greatest
impact on systems having highest throughput & memory requirements
High-Performance MPU (microprocessor unit) & ASIC (application-specific integrated circuit) Trends

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MPU/ASIC 1/2 pitch, nm</td>
<td>90</td>
<td>65</td>
<td>45</td>
<td>32</td>
<td>22</td>
</tr>
<tr>
<td>Transistors/chip</td>
<td>553M</td>
<td>1106M</td>
<td>2212M</td>
<td>4424M</td>
<td>8848M</td>
</tr>
<tr>
<td>Max watts @ volts</td>
<td>158@1.2</td>
<td>189@1.1</td>
<td>218@1.0</td>
<td>251@0.9</td>
<td><a href="mailto:288@0.8V">288@0.8V</a></td>
</tr>
<tr>
<td>Clock freq, MHz</td>
<td>4,171</td>
<td>9,285</td>
<td>15,079</td>
<td>22,980</td>
<td>39,683</td>
</tr>
<tr>
<td>Clock freq, MHz, for 158W power</td>
<td>4,171</td>
<td>7,762</td>
<td>10,929</td>
<td>14,465</td>
<td>21,771</td>
</tr>
</tbody>
</table>

- **2003 International Technology Roadmap for Semiconductors**
  - [http://public.itrs.net](http://public.itrs.net)
  - Executive summary tables 1i&j, 4c&d, 6a&b
  - Constant 310 mm² die size

- **Lithography improvement rate (partially driven by economics)** allows 2X transistors/chip every 3 yrs
  - 1.5X speed @ constant power
  - ~3X throughput for multiple independent ASIC (or FPGA) cores while maintaining constant power dissipation
  - ~2X throughput for large-cache MPUs (constant throughput/memory), but power may possibly decrease with careful design
Bottleneck Issues

- Bottlenecks occur when interconnection bandwidth (e.g., processor-to-memory, bisection or system-level I/O) is inadequate to support the throughput for a given application.

- For embedded applications, I/O bottlenecks are a greater concern for general-purpose, highly interconnected back-end vs. special-purpose, channelized front-end processors.

Can developers provide timely, cost-effective solutions to bottleneck problems?
Processor Bottlenecks at Device & System Levels

• Device level (ITRS03)
  – 2X transistors & 1.5X speed every 3 yrs
    High-performance microprocessor units & ASICs
    Constant power & 310 mm² die size
  – 3X throughput every 3 yrs possible if chip is mostly logic gates changing state frequently (independent ASIC or FPGA cores)
  – 2X throughput every 3 yrs is limit for microprocessors with large on-chip cache (chip is mostly SRAM & throughput/memory remains constant)
  – Possible technical solutions for microprocessors: 3D structures, on-chip controller for external L3 cache

• System level
  – 54W budget for hypothetical 6U COTS card computing 32 bit floating-point 1K complex FFT every 1µsec
    10% (5W) DC-to-DC converter loss
    40% (22W) I/O (7 input & 7 output links @ 10 Gbits/sec & 1.5W ea., RF coax, 2004)
    50% (27W) processor (51 GFLOPS sustained) & memory (5 Gbytes)
  – Possible technical solutions for I/O
    RF coax point-to-point serial links with central crosspoint switch network (PIN diodes or MEMS switches)
    Fiber optic links (may require optical free-space crosspoint switch) & optical chip-to-chip interconnects
Examples of Hardware vs. Algorithms

• Static RAM-based FPGAs
  – 2002: system-level throughput improved substantially vs. 1999
  – 2/3 of improvement attributable to new devices, 1/3 to architecture changes

• Chess computers
  – 1997: Deep Blue provided 40 trillion operations per second using 600nm custom ASICs (but 250nm was state-of-the-art)
  – 2001: Desktop version of Deep Blue using state-of-the-art custom ASICs feasible, but not built
  – 2002-2003: improved algorithms provide functional equivalent of Deep Blue using COTS servers instead of custom ASICs

• Speedup provided by FFT & other “fast” algorithms

Contributions of HW vs. algorithms may be difficult to quantify, even when all necessary data are available
Initial operating capability SW has 12% HW utilization, allowing 8X growth over 9 yr lifetime (2X every 3 yrs): HW still programmable @ end-of-life.

Frequent SW-only “tech refresh” provides upgraded capabilities for fixed HW in satellites & space probes, ship-based missiles & torpedoes, radars, “software radios,” etc.
Timeline for Highest Performance COTS ADCs, 2Q04

- 2X speed (up to 1/2 bit processing gain) in ~4.5 yrs for high-speed ADCs
- 6- to 8-bit (ENOB>5.0)
- 10-bit (ENOB>7.6)
- 12-bit (ENOB>9.8)
- 14-bit (ENOB>11.8)
- 16-bit (ENOB>12.5)
- 24-bit (ENOB>15.3)

Dashed lines indicate upper bounds (all ADCs below a dashed line have Effective Number Of Bits indicated)
Improvement Rates for
Highest Performance COTS ADCs, 2Q04

ADC improvements @ ~100 KSPS
limited by commercial audio market

1/2 bit/octave maximum processing gain with linearization

~1/3 bit/yr for high-resolution ADCs
@ 100 MSPS

~1 bit/octave slope

ADC improvements @ ~100 KSPS
limited by commercial audio market

2X speed (up to 1/2 bit processing gain)
in ~4.5 yrs for high-speed ADCs

Effective Number of Bits

0.1 1 10 100 1000 10000

Sampling Rate (million samples/sec)
Evolution of COTS Embedded Multiprocessor Cards, 2Q04

- **Computation Efficiency (GFLOPS/Watt)**
- **Computation Density (GFLOPS/Liter)**

**GFLOPS** (billions of 32-bit floating point operations/sec) sustained for 1K complex FFT

- **Reconfigurable FPGA cards (~100 FLOPS/byte)** improving 3X in 3 yrs
- **Special-purpose ASIC cards (~10 FLOPS/byte)** improving 3X in 3 yrs
- **General-purpose RISC (with on-chip vector processor) cards (~10 FLOPS/byte)** improving 2X in 3 yrs

**71 W/Liter limit for convection-cooled cards**

- **3/99**: 7/99
- **3/00**: 12/03
- **6/10**: 10/04
- **7/05**: 11/01
- **8/02**: 12/03
- **6/10**: 10/04

MIT Lincoln Laboratory
Timeline for Highest Performance COTS Multiprocessors, 2Q04

Open systems architecture goal: mix old & new general- & special-purpose cards, with upgrades as needed (a new card may replace four 3-yr-old cards)
Timeline for COTS Processor I&O Rate and ADC Sampling Rate (2Q04)

- Effective number of bits @ sampling rate for high-speed, high-resolution analog-to-digital converters

- Card-level I&O complex sample rate sustained for 32 bit flt-pt 1K complex FFT (1000 MSPS for 50 GFLOPS)

- Open systems architecture goal: mix old & new general- & special-purpose cards, with upgrades as needed (a new card may replace four 3-yr-old cards)

- Moore's Law slope: 4X in 3 yrs

- Field-programmable gate arrays

- Future FPGA

- Future ASIC

- Future ADC

- i860 µP

- SHARC DSP

- Pathfinder-1 ASIC

- Virtex II FPGA

- TM-44 Blackbird ASIC

- CS301 ASIC

- PowerPC RISC with AltiVec

- Virtex FPGA

- Virtex FPGA

- Virtex FPGA

- Future FPGA

- Future ASIC

- Future ADC

- Special-purpose application-specific integrated circuits

- Programmable microprocessors, digital signal processors & reduced instruction set computers

- Highest-performance 6U form factor multiprocessor cards <55W