Exceptional service in the national interest

Sandia National Laboratories

Why Reversible Computing is the Only Long-Term Path for Sustained, Affordable Performance Growth

Michael P. Frank
Center for Computing Research
Sandia National Laboratories

Presented at the Computational Science Seminar, Sandia
January 31, 2017, Albuquerque, NM

Approved for Unclassified Unlimited Release
SAND2017-1657 PE

U.S. DEPARTMENT OF ENERGY
NNSA National Nuclear Security Administration

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

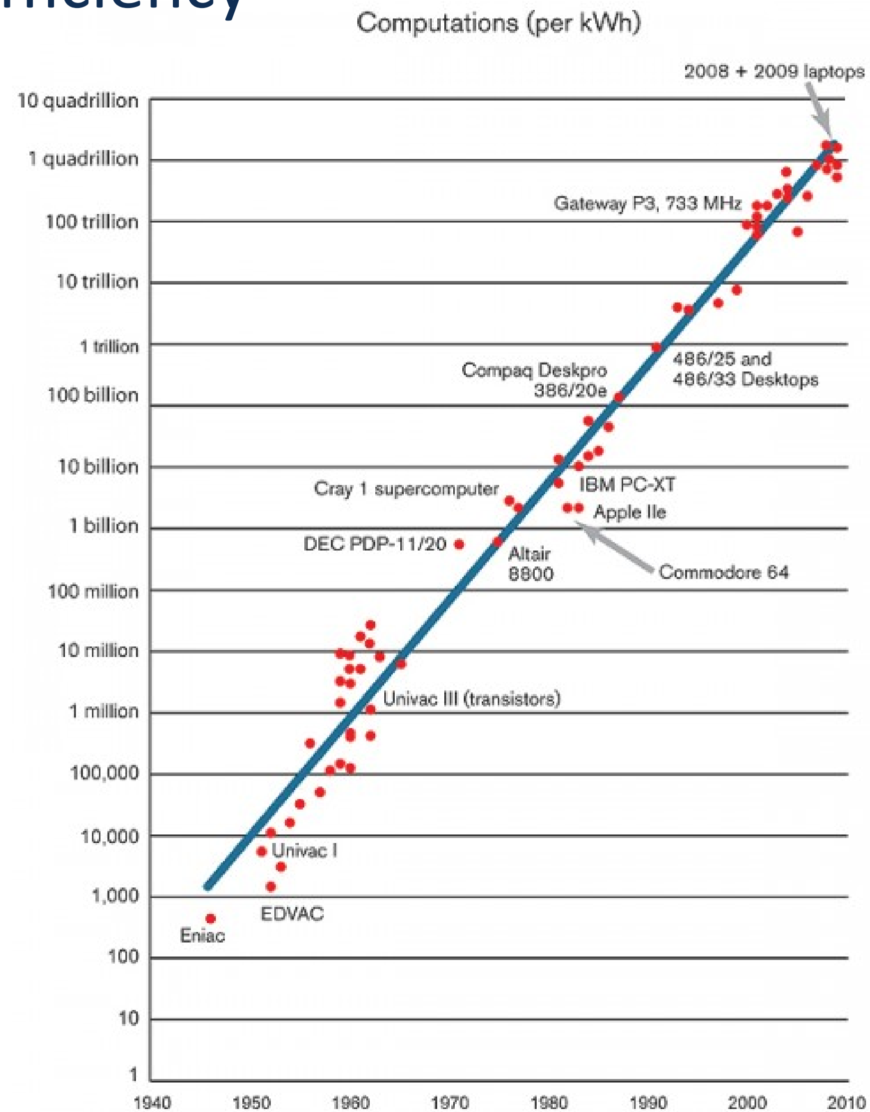
Structure of the Talk

1. The Historical Power-Performance Trend
   - A key engine of economic growth, which must not stall
2. The Thermodynamic Brick Wall of Irreversible Computing
   - Why it truly is absolutely unavoidable, *except* by reversible computing
3. Reversible Computing Theory – Basic Concepts
   - Limitations of the classic models, and how to fix them.
4. Progress Towards Practicality
   - Gradual improvements in implementation concepts
5. The Challenges Yet to be Faced
   - What are the hard problems in RC that still need to be solved?
6. Conclusion

The Power-Performance Trend
and the importance of energy efficiency

Sandia National Laboratories

- *Any* system (at any scale) scoped to have a fixed cost-of-ownership over its operational lifetime *must* implicitly carry some associated maximum budget for all energy-related costs.
  - These costs include things like:
    - In mobile devices, cost of batteries and inconvenience to user of charging
    - kWhr electricity costs for desktop owners
    - Cost to build and operate high-capacity machine room/datacenter AC systems
    - Cost to build or lease a nearby power plant if required to supply an exascale machine
- We can't expect the cost of <u>energy</u> to ever decrease by orders of magnitude.
  - Essentially, energy *is* "nature's currency."
- Thus, fundamentally, *increasing affordable performance requires increasing computational energy efficiency*. (Useful ops done/Joule.)
  - And this has, indeed, been the historical trend, for >50 years.

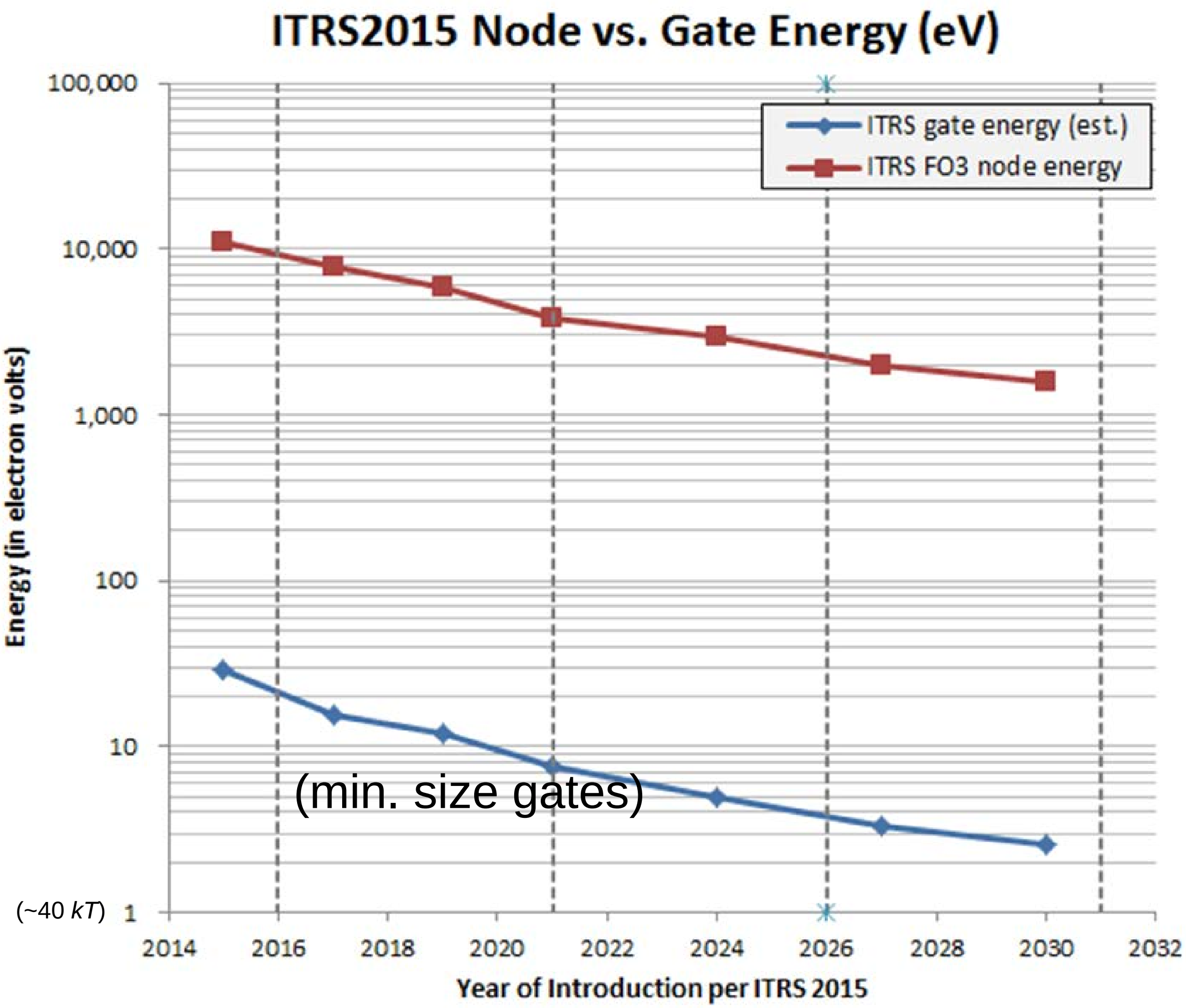


(MIT Technology Review, Apr. 2012)

Energy limits for conventional technology are not that far away!

Sandia National Laboratories

- Energy of min.-width FET gates affects channel fluctuations < ~1-2 eV
  - Impact on leakage
- Real gates are often wider (~ 20x min.)
  - Also there is wire / junction capacitance
- Note: ITRS is aware of thermal noise issue, and so has min. gate energy asymptoting to ~2 eV
  - Node energy *follows*, asymptoting to ~1 keV
- Practical circuit architectures can't just magically cross this gap!
  - ∴ Fundamental thermal limits translate to much *larger* practical limits!



Fundamental Thermal Limits on all Conventional (Irreversible!) computing

- Limits due to thermal noise:
  - Due to the fundamental arguments for the Boltzmann distribution,
    - to suppress the probability or rate of thermally-induced transitions to/through undesired states by a factor of R requires an energy difference ΔE between desired and undesired states of $\Delta E \cong k_{\mathrm{B}} T \ln R$.
  - In conventional logic schemes, this energy difference translates (together w. overheads of prev. slide) into a minimum logic signal energy,
    - which is dissipated to heat every time a node's logic value is cycled.
      - But, there are other *unconventional* schemes in which the logic signal energy can itself be even less than ΔE, while still maintaining reliable overall operation
        - See my "Chaotic Logic" talk, ICRC 2016
      - Moreover, even when the signal energy is large, *this energy does not need to be dissipated to heat in order to do useful logic with it*!
        - Recovery and reuse of an amount of energy approaching the *entire* signal energy is possible using reversible logic!
- Fundamental information-theoretic limit (Landauer's principle)
  - Very simple, irrefutable limit! (See next few slides)

Information Loss = Entropy Increase

- All <u>fundamental</u> physical dynamics is (microscopically) *reversible*.
  - Any Hamiltonian dynamical system:
    - Let the time increments δt be negative → Time-evolution runs in reverse.
  - Quantum mechanical time-evolution (generalized Schrödinger equation):
    - Any two quantum states that are initially mutually distinguishable (orthogonal) will always remain so, under any unitary time-evolution operator, $U(t) = \mathrm{e}^{-iHt/\hbar}$.
- ∴ *Detailed physical information can never, ever be destroyed!*
  - Only reversibly transformed, in place (locally)!
    - At most, we can only *lose track* (from a modeling perspective) of the (always-still-microscopically-reversible) transformations that have occurred.
      - Uncertainty increase → Effective randomization of the detailed state
  - If this were not true, the 2nd Law of Thermodynamics would not hold!
    - Effectively, entropy is simply that portion of the total physical information that happens to have already been randomized/scrambled beyond any hope of practically transforming it back into its original form.
      - ∴ If information could be destroyed, then entropy could simply vanish
- *To "irreversibly lose information"* <u>means</u> for that information to be (reversibly) transformed in any way that we cannot <u>*practically*</u> undo.
  - It's "lost" in the sense that its original form cannot be <u>practically</u> recovered.
  - "Irreversible information loss" <u>is exactly the same thing as</u> "entropy increase."

Landauer's Principle—
A Simplified Statement:

- For each bit's worth of local information that is irreversibly lost from (*e.g.*, obliviously "erased" by , or "destructively overwritten" by) any computational device encompassed by a thermal environment at temperature T, no less than an amount

$$E_{\text{diss}} = k_B T \ln 2$$

  of free energy ("Landauer's limit") must eventually be dissipated as heat added to that thermal environment.
  - This is easily proven, as a theorem of applied mathematical physics.
- *Approachability hypothesis:*
  - Landauer's bound may be approached arbitrarily closely in a suitably-designed family of realistically-constructible physical mechanisms.
    - Abstract physical procedures described in the literature support this.

Landauer's Principle—
A Correct *General* Formulation:

- Consider any computational device D that is designed to transform initial logical states $s_{\mathrm{I}} \in S_{\mathrm{I}} = \{s_{\mathrm{I}1}, s_{\mathrm{I}2}, \dots, s_{\mathrm{I}n}\}$ to final logical states $s_{\mathrm{F}} \in S_{\mathrm{F}} = \{s_{\mathrm{F}1}, s_{\mathrm{F}2}, \dots, s_{\mathrm{F}m}\}$ according to some (in general probabilistic) transition rule, $r_i(j) = \Pr\left[s_{\mathrm{F}} = s_{\mathrm{F}j} \middle| s_{\mathrm{I}} = s_{\mathrm{I}i}\right]$.
  - Now consider any given probability distribution over initial states, $p_{\mathrm{I}}(i) = \Pr[s_{\mathrm{I}} = s_{\mathrm{I}i}]$, defining a given statistical scenario in which D is to be operated. (An "*operation context.*")
    - The entropy $H[p_{\mathrm{I}}]$ of this initial state distribution is:

$$H[p_{\mathrm{I}}] = \sum_{i=1}^{n} p_{\mathrm{I}}(i) \ln \frac{1}{p_{\mathrm{I}}(i)}.$$

    - And, after D has operated, we can derive, from p_{I} and $r_i(j)$, the final state distribution p_F, which is

$$p_{\mathrm{F}}(j) = \Pr\left[s_{\mathrm{F}} = s_{\mathrm{F}j}\right] = \sum_{i=1}^{n} p_{\mathrm{I}}(i) \cdot r_i(j).$$

    - And the entropy $H[p_{\mathrm{F}}]$of the final state distribution is:

$$H[p_{\mathrm{F}}] = \sum_{j=1}^{m} p_{\mathrm{F}}(j) \ln \frac{1}{p_{\mathrm{F}}(j)}.$$

    - Then, the minimum entropy ejected from the device D as a side-effect of its operation in context p_{I} must be:

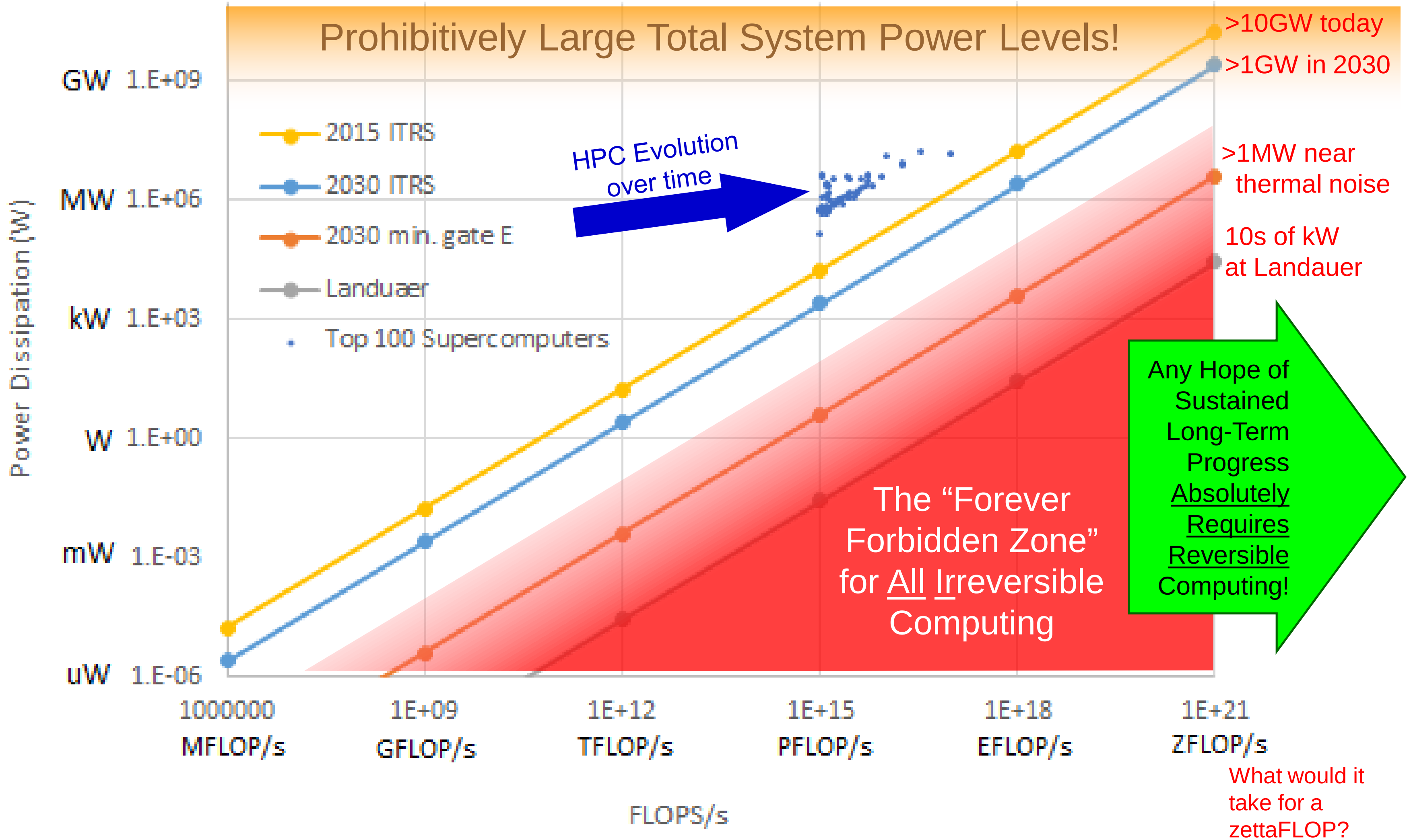
$$\Delta H_D(p_{\mathrm{I}}) = H[p_{\mathrm{I}}] - H[p_{\mathrm{F}}],$$

    since total entropy cannot decrease (by fundamental reversibility/the 2nd law of thermodynamics).

- Therefore, device D, when operated in a statistical context p_{I}, necessarily loses an amount of information (*i.e.*, ejects an amount of entropy) $\Delta H_D(p_{\mathrm{I}})$.
  - Suppose this entropy eventually ends up in some external thermal reservoir at temperature T.
  - Then, by the thermodynamic definition of temperature, we must add heat $\Delta Q = T\Delta H_D(p_{\mathrm{I}})$ to the reservoir.

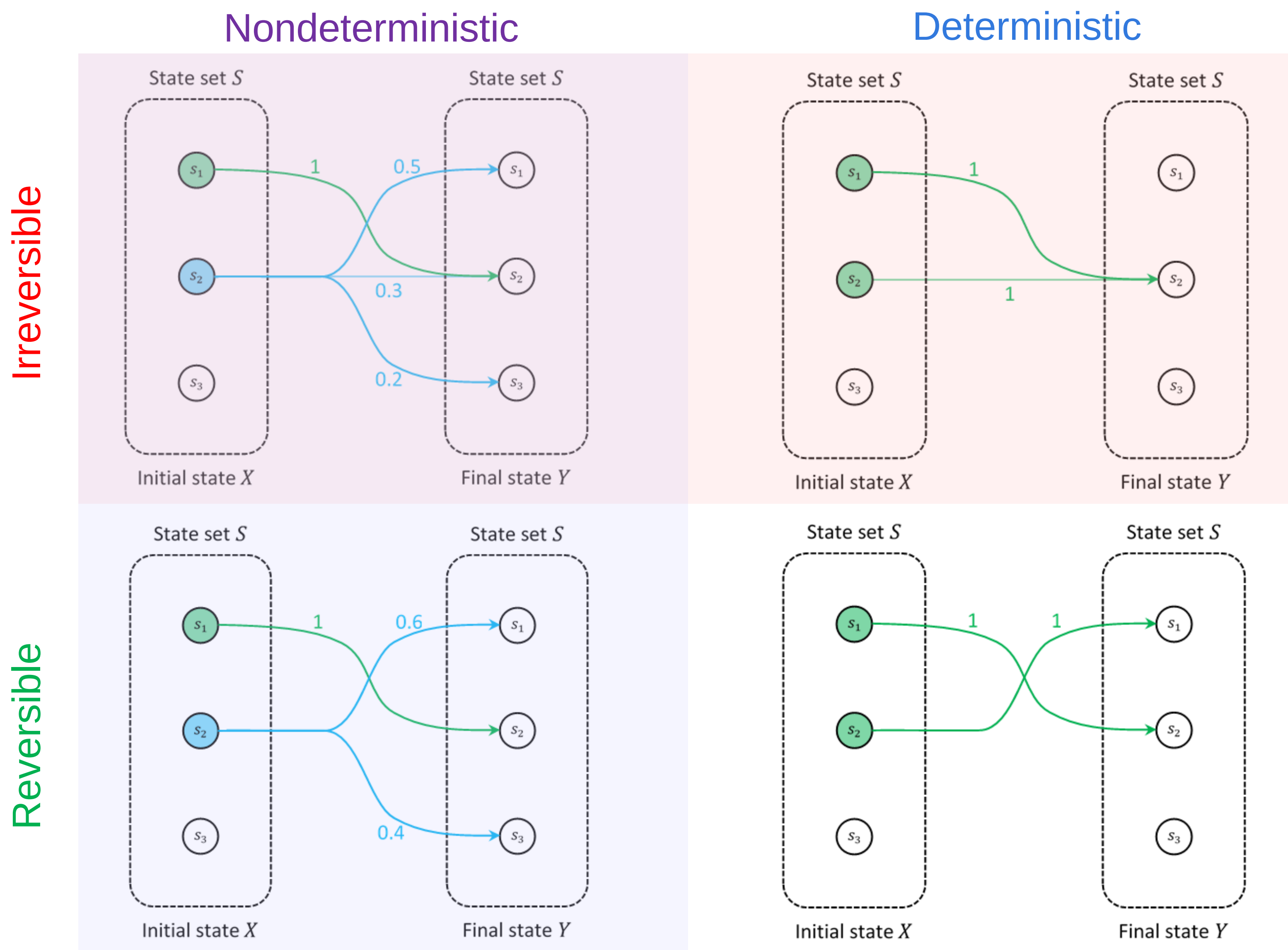
Implications for FLOPS & power

Note: The limits suggested by the diagonal lines do not even include power for interconnects, memory, or cooling!



Types of Computational Operations

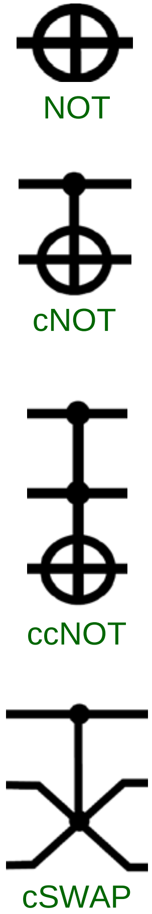
Define operations as (possibly partial) probabilistic transition relations



Unconditionally Reversible (UR) Gates

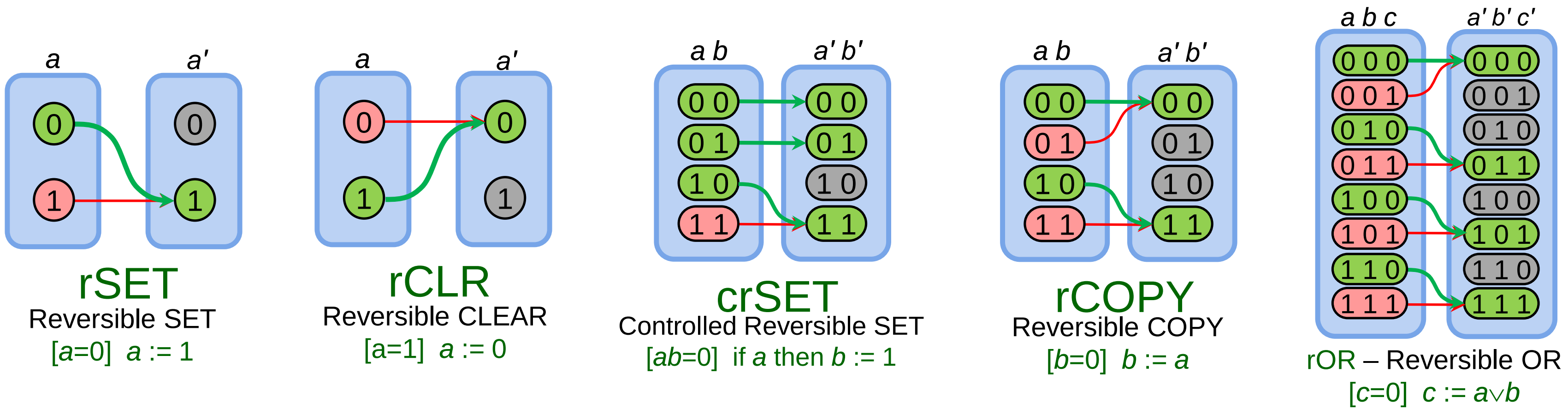
(These are only a special case!)

- Any total, reversible, deterministic operation is simply a permutation (bijective transformation) of the state set.
- Some example UR operations (misleadingly called "gates") on binary-encoded states:
  - NOT(*a*) — *a* := ¬*a* — In-place bit-flip
  - cNOT(*a*,*b*) — if *a*=1 then *b* := ¬*b* — Controlled NOT
  - ccNOT(*a*,*b*,*c*) — if *ab*=1 then *c* := ¬*c* — A.k.a. "Toffoli gate"
  - cSWAP(*a*,*b*,*c*) — if *a*=1 then *b* ↔ *c* — A.k.a. "Fredkin gate"
- ccNOT and cSWAP are each universal UR gates
  - The latter in the case of functions on dual-rail-encoded bit-strings
- No set of just 1- and 2-bit classical UR gates is universal
  - However, cNOT plus 1-bit quantum (unitary) gates comprise a universal set



Generalized Reversible Computing (GRC) also includes Conditional Reversibility (CR)!

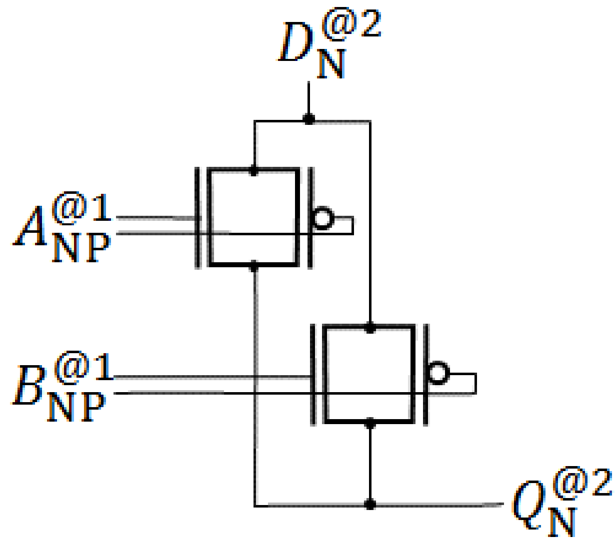
- Definition: A (deterministic) operation O is *conditionally reversible under precondition* $P \subseteq S$ if and only if the restriction of O to P (as a partial operation) is an injective (one-to-one) operation.
  - Given any initial probability distribution p over states in S such that $p(x) = 0$ for all $x \notin P$, the application of the operation O does not reduce the entropy of the computational state at all, and so incurs no minimum dissipation under Landauer's principle.
    - And, as all those $p(x) \rightarrow 0$, so does the minimum Landauer dissipation.
- Examples of some conditionally reversible operations:
  - Green denotes the restriction of the operation to the precondition
  - Red: States that would result in dissipation b/c precondition not met

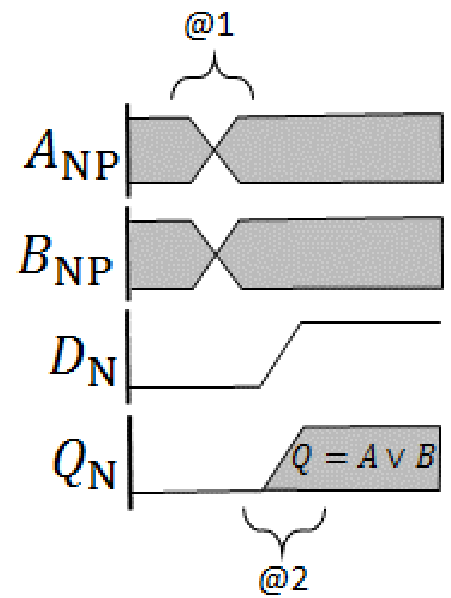


Implementing Conditionally-Reversible Operations

Sandia National Laboratories

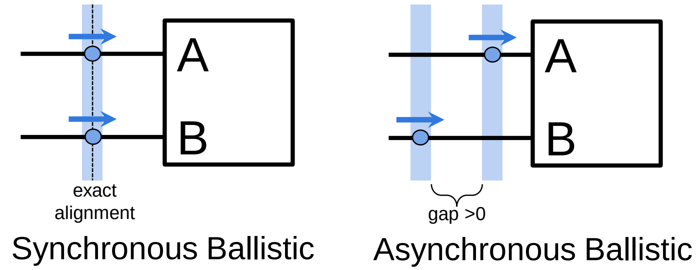
- Not very difficult!
  - Straightfoward to do with adiabatic switching
- E.g., this CMOS structure can be used to do/undo latched rOR operations
  - Example of 2LAL logic family
    - Based on CMOS transmission gates
    - Implicit dual-rail complementary signals (PN pairs) in this notation
- Computation sequence:
  1. Precondition: Output signal *Q* initially at logic `0`
  2. Driving signal *D* is also initially logic `0`
  3. At time 1 (@1), inputs *A*, *B* transition to new levels
     - Connecting *D* to *Q* if and only if *A* or *B* is logic `1`
  4. At time 2 (@2), driver *D* transitions from `0` to `1`
     - *Q* follows it to `1` if and only if *A* or *B* is logic `1`
     - Now *Q* is the logical OR of inputs *A*,*B*
- Reversible things that we can do afterwards:
  - Restore *A*, *B* to `0` (latching *Q*), or, undo above steps

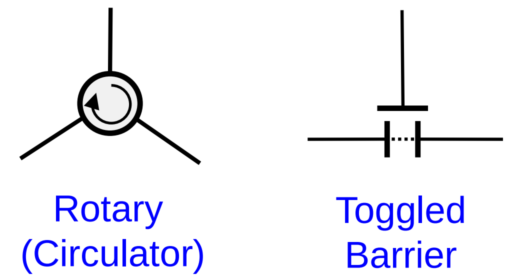




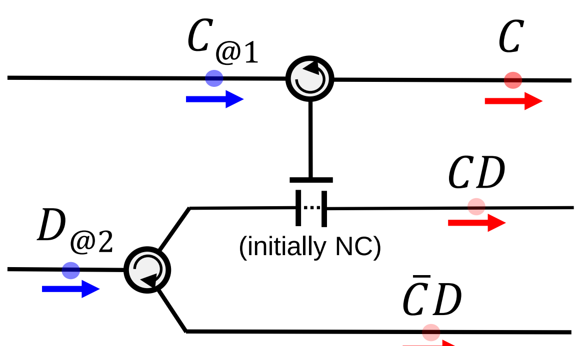
Asynchronous Ballistic Reversible Computing

- Some problems with all of the existing *adiabatic* schemes for reversible computing:
  - In general, numerous power/clock signals are needed to drive adiabatic logic transitions
  - Distributing these signals adds substantial complexity overheads and parasitic power losses
- Ballistic logic schemes can eliminate the clocks!
  - Devices simply operate whenever data pulses arrive
  - The operation energy is carried by the pulse itself
    - Most of the energy is preserved in outgoing pulses
    - Signal restoration can be carried out incrementally
- But, *synchronous* ballistic logic has some issues:
  - Unrealistically precise timing alignment required
  - Chaotic amplification of timing uncertainties when signals interact
- Benefits of asynchronous ballistic logic:
  - Much looser timing constraints
  - Linear instead of exponential increase in timing uncertainty per logic stage
  - Potentially simpler device designs
- New effort to investigate implementing ABRC in superconducting circuits (N&M LDRD idea)…





Example ABR device functions



Example logic construction

Scaling of Reversible Computation

- A significant tradeoff that comes into play in applying reversible computing is that reversible hardware designs typically incur some moderate overheads in terms of hardware complexity (per unit performance)…
  - Small polynomial overheads, as a function of the energy efficiency boost obtained. (Precise scaling depends on the problem class)
    - Typically at most only linear, or slightly more than linear overhead
- Despite these hardware overheads, there are two strong arguments as to why reversible computing still stands as the dominant long-term path forwards (next two slides):
  - Fundamental economic argument
  - Fundamental physics of computing argument

Fundamental Economic Argument

- The *ultimate* measure of cost is always energy ("nature's currency"), or (more precisely) negentropy
  - Even manufacturing costs ultimately derive from the energy used to mine/refine/assemble materials, feed members of the workforce, *etc.*
- However, we know no fundamental reasons why per-device manufacturing costs cannot become arbitrarily close to 0, through ongoing manufacturing process innovations…
  - In the distant future, we can even imagine doing "reversible manufacturing," in which materials are rearranged via thermodynamically reversible nanoscale manipulations of individual atoms
- Meanwhile, doing *more computation* enables delivery of *more economic value* in general (we assume)
  - Therefore, in the long run, being able to carry out an ever-increasing number of useful operations, per Joule of energy dissipated, can easily pay for the correspondingly increased hardware complexity, as per-device manufacturing costs continue to decrease.

Fundamental Physics of Computing Argument

- Since the underlying physics is itself always reversible, computers that are based on traditional irreversible computing design principles are really <u>only a special case</u>…
  - One in which we are restricting ourselves to a limited subset of designs, those in which our method of handling garbage information is to always just treat it as entropy and move it out of the machine
- The more *general* design space, which includes designs capable of utilizing reversible computing, and decomputing some of the garbage rather than expelling it, cannot possibly be any *worse* than the *limited* irreversible design space…
  - And it's possible to prove that, given any fixed finite constraint on heat flux density, reversible machines asymptotically scale strictly *better*, *<u>even</u>* when we ignore the cost of energy (*c.f.* my dissertation)
    - Because denser packing of components → lower communication delays

Some Highlights of Reversible Computing History

- 1961 – Landauer's original paper on thermal cost of irreversibility
- 1973 – Bennett, *Logical Reversibility of Computing*
- Late 1980s – Feynman, Margolus –
  - Quantum-mechanical models of reversible computing
- 1989 – Bennett, more space-efficient reversible algorithms
- 1980s, 1990s – Various groups
  - Early adiabatic MOS-based circuits, various alternative implementation proposals
    - superconducting, nanomechanical, quantum dot based, *etc.*
- Late 1990s/early 2000s – Myself and others
  - Reversible computer architectures, scaling analyses
- 2009-present – Progress in various CS theory aspects
  - Annual conference on reversible computation, several books
- Also progress in adiabatic & superconducting implementations…

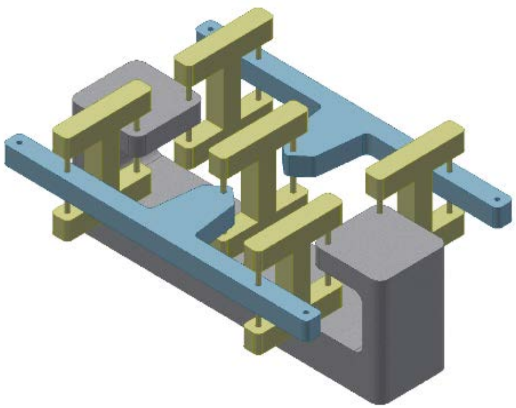
Key Challenges for the Field

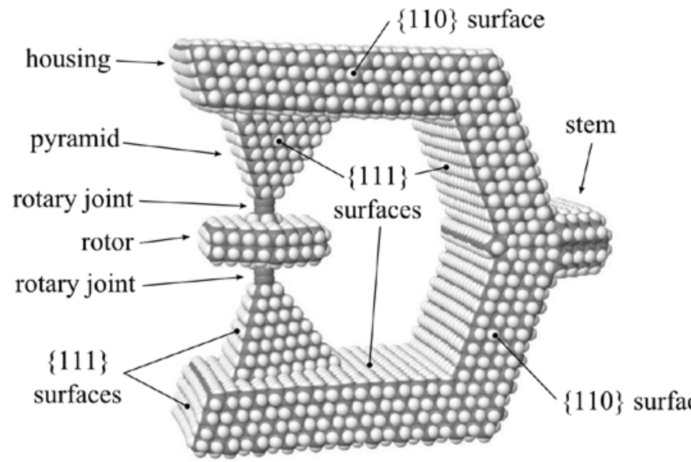
- Develop new manufacturable device technologies offering improved performance characteristics for reversible operation
  - One key goal: Low adiabatic energy coefficient, $c_{\mathrm{E}} = E_{\mathrm{diss}} \cdot t_{\mathrm{op}}$
    - For cryogenic technologies, adjust this to account for cooling overheads
  - New devices facilitating ABRC would be very desirable
  - Per-device manufacturing cost is, of course, also still important
- Develop new logic models, logic circuit architectural styles, hardware algorithms, *etc.* that can utilize the new devices
  - *E.g.*, GRC model in general, 2LAL logic family for adiabatic CMOS, ABRC model for pulse-based (e.g., SFQ) ballistic logics
  - There is a significant literature now addressing reversible algorithms
    - A few books, an annual conference on reversible computation
- Significant new investments in tool development are needed:
  - *E.g.*, EDA tools, hardware description languages
  - Eventually: Reversibility-aware programming languages/compilers

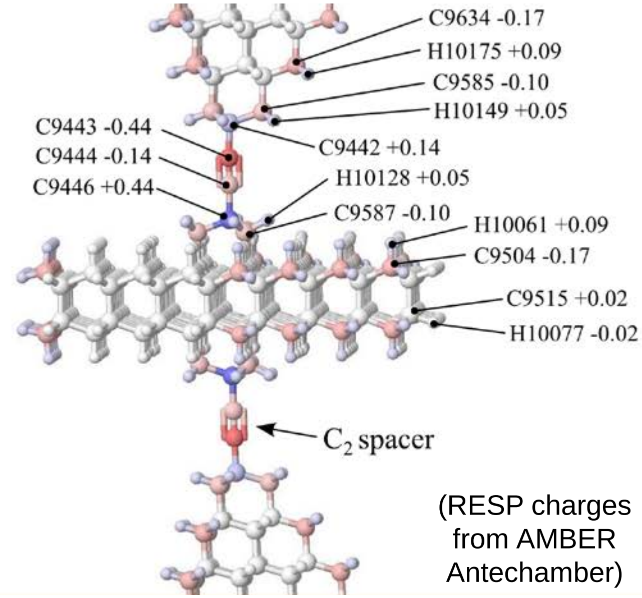
Nanomechanical Rotary Logic

Sandia National Laboratories

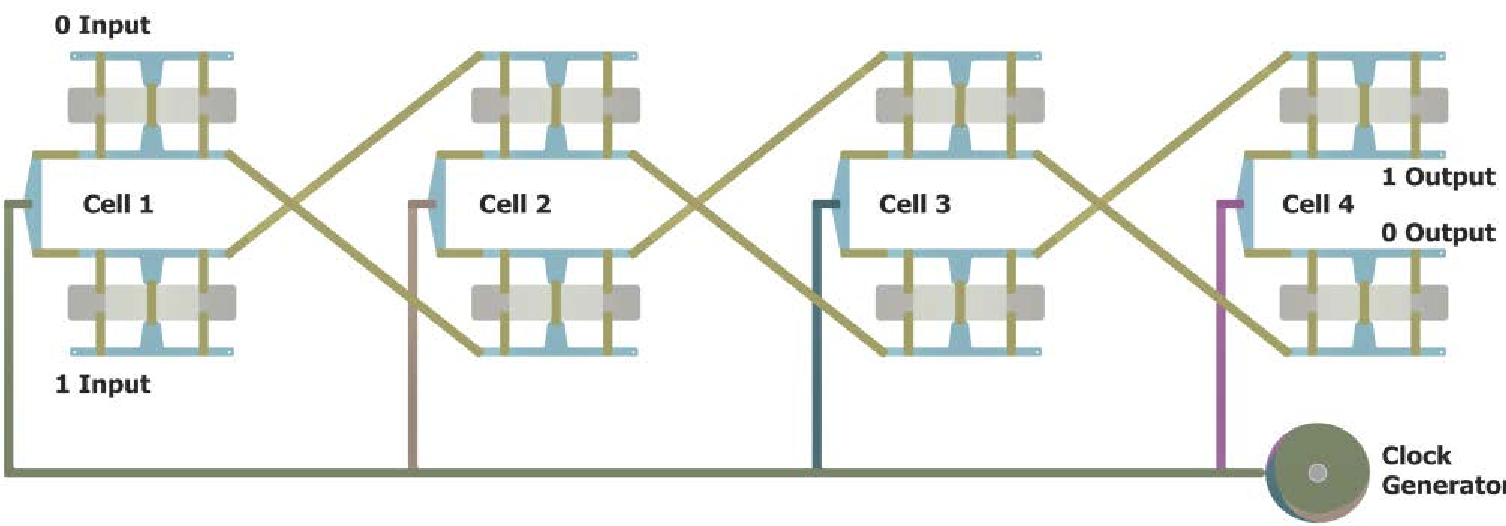
Merkle et al., IMM Report 46 and Hogg et al., arxiv:1701.08202 (reproduced with permission)





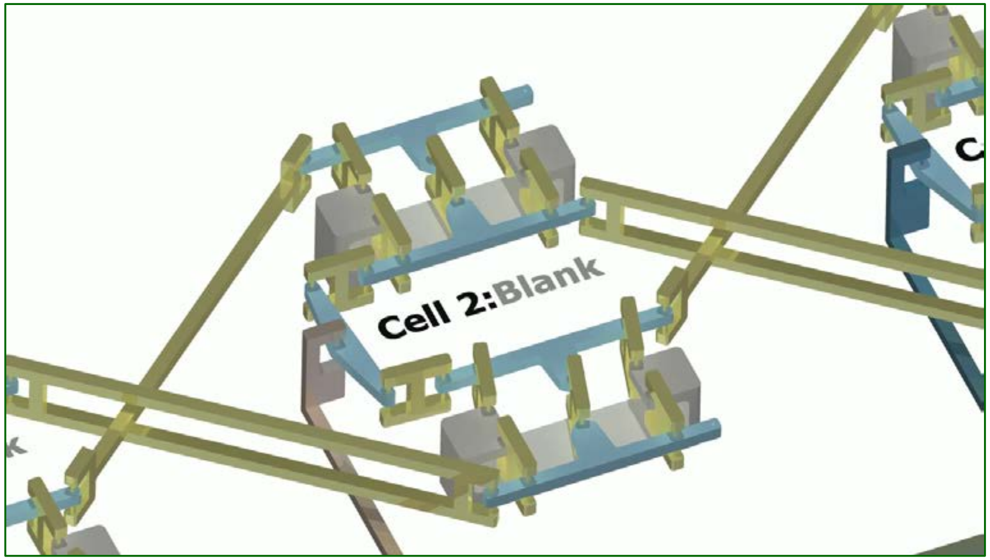


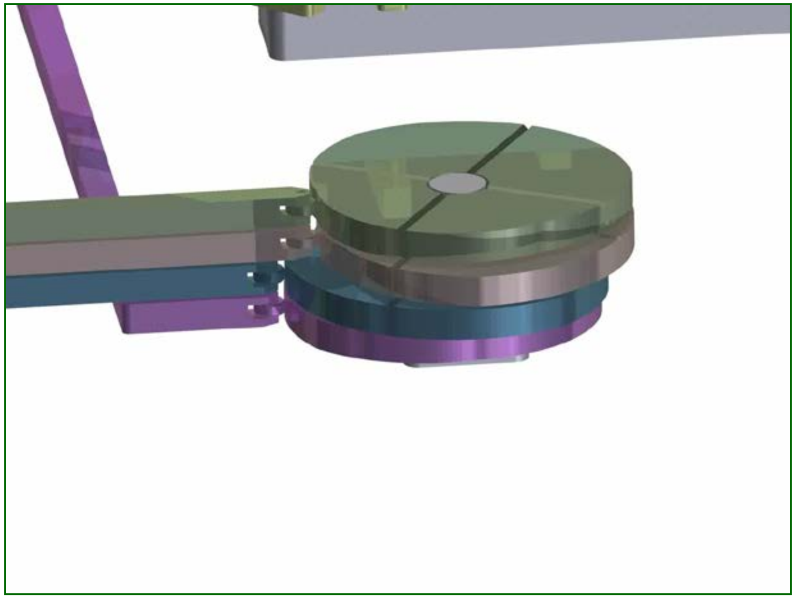
(RESP charges from AMBER Antechamber)



Rotary Logic Lock Operation

- Videos animate schematic geometry of a pair of locks in a shift register
- Molecular Dynamics modeling/simulation tools used for analysis include:
  - LAMMPS, GROMACS, AMBER Antechamber
- Simulated dissipation:
  - ~4 × 10^{-26} J/cycle at 100 MHz
    - 74,000 × lower than the Landauer limit for irreversible ops!
- Speeds up into GHz range should also be achievable





Conclusion

- The computer industry is facing imminent thermodynamic roadblocks, which will soon prevent very much further progress in practical performance/cost…
  - Any physically possible general-purpose irreversible computing technology can be expected to face practical energy efficiency barriers at roughly the same order of magnitude as for end-of-roadmap CMOS (within 10-100x, most likely)
- The *only* physically possible *general-purpose* solution that has the potential to sustain affordable performance growth over *many* technology generations (and not just a few) is to use some form of *reversible computing*…
  - Quantum computing is also great, if it can be done, but its applicability is more limited…
  - Analog/neuromorphic approaches are subject to the same laws of thermodynamics!
    - They, too, can only continue increasing in energy-efficiency if *they are also reversible!*
- Traditional theoretical models of reversible logic are unnecessarily restrictive…
  - The concepts of *conditional reversibility* and *asynchronous reversible computing* illustrate useful ways of generalizing them, to facilitate practical hardware design
- New, much more efficient reversible device technologies are badly needed…
  - But, creative new implementation concepts (such as Rotary Logic) illustrate that there is no fundamental physical reason why such improved technologies cannot exist!
- If we don't want progress to stall, reversible computing *must* be developed to the point where it can take over as the overwhelmingly dominant foundational paradigm for most general-purpose computing looking forward…
  - It's high time we begin serious new R&D efforts to make this happen!