MDGRAPE-4 NoC Links

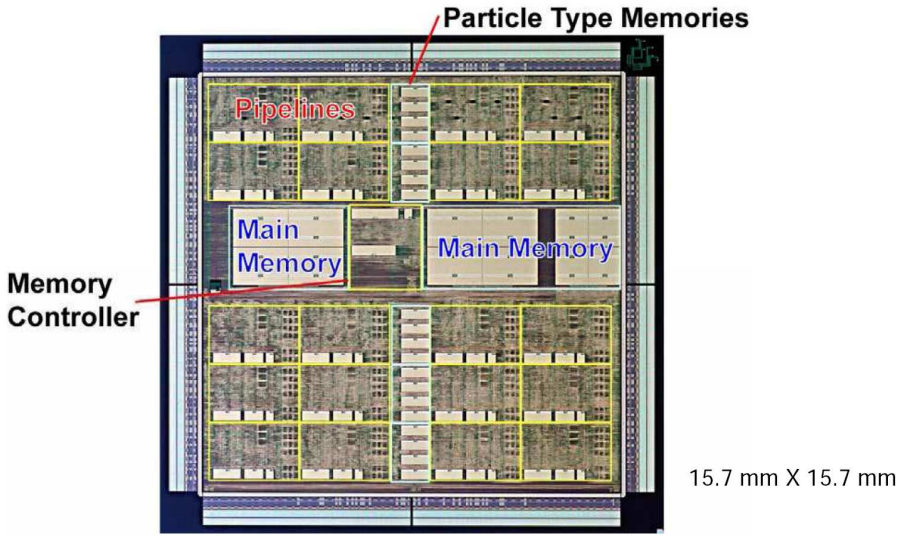
Circuit and Physical Design

Duraid Madina, Makoto Taiji

SLIP 2008

MDGRAPE-3

MDGRAPE-3 is an efficient chip that accelerates n-body simulations:



Vital stats:

- 130nm, 1.2V, 7Cu, 250mm^2
- 230GFLOPS @ 19W, 350MHz (250MHz worst-case)
- 20 force calculation pipelines (˜single precision): ˜660 flops/cycle

MDGRAPE-3

Put 12 chips on a board:

(˜2.7TFLOPs, ˜250W)

MDGRAPE-3

Put 400 boards in a room, sprinkle Xeon boxes:

(~1PFLOPs, ~200kW, 22 racks - that's all of them)

MDGRAPE-3

MDGRAPE-3:

- Works great (first running petascale machine)

MDGRAPE-3

MDGRAPE-3:

- Works great (first running petascale machine)
- Reasonable cost
  - Chip design: one person + some foundry help
  - 200kW: normal room, normal aircon, no hearing loss

MDGRAPE-3

MDGRAPE-3:

- Works great (first running petascale machine)
- Reasonable cost
  - Chip design: one person + some foundry help
  - 200kW: normal room, normal aircon, no hearing loss
- However...
  - Single purpose (more or less)

MDGRAPE-4

Work towards MDGRAPE-4 is well underway. We're aiming at:

- A broader range of scientific tasks
  - A heterogenous, tiled processor
  - Lightweight processor cores and special-purpose units

MDGRAPE-4

Work towards MDGRAPE-4 is well underway. We're aiming at:

- A broader range of scientific tasks
  - A heterogenous, tiled processor
  - Lightweight processor cores and special-purpose units
- A more efficient design
  - Moving from limited use of custom cells to more widespread custom blocks
  - Low power design used wherever possible to enable high system performance through scalability

MDGRAPE-4 Early Design

- Limited resources further motivated tiled architecture

MDGRAPE-4 Early Design

- Limited resources further motivated tiled architecture
  - If we can get some tile designs finished, we're done.

MDGRAPE-4 Early Design

- Limited resources further motivated tiled architecture
  - If we can get some tile designs finished, we're done.
- Quick-and-dirty ports of existing blocks to the new process and back-of-envelope numbers crunched to determine three key figures:

MDGRAPE-4 Early Design

- Limited resources further motivated tiled architecture
  - If we can get some tile designs finished, we're done.
- Quick-and-dirty ports of existing blocks to the new process and back-of-envelope numbers crunched to determine three key figures:
  - Total throughput

MDGRAPE-4 Early Design

- Limited resources further motivated tiled architecture
  - If we can get some tile designs finished, we're done.
- Quick-and-dirty ports of existing blocks to the new process and back-of-envelope numbers crunched to determine three key figures:
  - Total throughput
  - Power consumption

MDGRAPE-4 Early Design

- Limited resources further motivated tiled architecture
  - If we can get some tile designs finished, we're done.
- Quick-and-dirty ports of existing blocks to the new process and back-of-envelope numbers crunched to determine three key figures:
  - Total throughput
  - Power consumption
  - Approximate tile sizes

MDGRAPE-4 Floorplanning

- Estimated tile sizes of $1mm^2$

MDGRAPE-4 Floorplanning

- Estimated tile sizes of $1mm^2$
- Significant local traffic flow (e.g. pipelined force accumulation)

MDGRAPE-4 Floorplanning

- Estimated tile sizes of $1mm^2$
- Significant local traffic flow (e.g. pipelined force accumulation)
- However, didn't want to hit Amdahl's law too early, so identified folded torus-like network topology as offering reasonable balance:

MDGRAPE-4 Floorplanning

- Estimated tile sizes of $1mm^2$
- Significant local traffic flow (e.g. pipelined force accumulation)
- However, didn't want to hit Amdahl's law too early, so identified folded torus-like network topology as offering reasonable balance:
  - Good performance under heavy local traffic

MDGRAPE-4 Floorplanning

- Estimated tile sizes of $1mm^2$
- Significant local traffic flow (e.g. pipelined force accumulation)
- However, didn't want to hit Amdahl's law too early, so identified folded torus-like network topology as offering reasonable balance:
  - Good performance under heavy local traffic
  - Improved throughput under difficult traffic, without cost (or performance) of hypercube/fat-tree networks

MDGRAPE-4 Floorplanning

- Estimated tile sizes of $1mm^2$
- Significant local traffic flow (e.g. pipelined force accumulation)
- However, didn't want to hit Amdahl's law too early, so identified folded torus-like network topology as offering reasonable balance:
  - Good performance under heavy local traffic
  - Improved throughput under difficult traffic, without cost (or performance) of hypercube/fat-tree networks
  - Makes use of longer (˜2.2mm, ˜4.5mm) links for low latency comms in time-critical parts of code

MDGRAPE-4 Floorplanning

- Estimated tile sizes of 1mm^2
- Significant local traffic flow (e.g. pipelined force accumulation)
- However, didn't want to hit Amdahl's law too early, so identified folded torus-like network topology as offering reasonable balance:
  - Good performance under heavy local traffic
  - Improved throughput under difficult traffic, without cost (or performance) of hypercube/fat-tree networks
  - Makes use of longer (˜2.2mm, ˜4.5mm) links for low latency comms in time-critical parts of code
  - Makes good use of process' semi-global layers

MDGRAPE-4 Floorplanning

- Estimated tile sizes of $1mm^2$
- Significant local traffic flow (e.g. pipelined force accumulation)
- However, didn't want to hit Amdahl's law too early, so identified folded torus-like network topology as offering reasonable balance:
  - Good performance under heavy local traffic
  - Improved throughput under difficult traffic, without cost (or performance) of hypercube/fat-tree networks
  - Makes use of longer (˜2.2mm, ˜4.5mm) links for low latency comms in time-critical parts of code
  - Makes good use of process' semi-global layers
  - Upper local layers freed for intra-tile and neighbouring-tile full-swing buses.

MDGRAPE-4 Floorplanning

- Prototyped a number of bus layouts using a periodic router

MDGRAPE-4 Floorplanning

- Prototyped a number of bus layouts using a periodic router
  - Very simple (100s of LoC) A* router with one distinguishing feature: support for per net/per bus periodic boundaries

MDGRAPE-4 Floorplanning

- Prototyped a number of bus layouts using a periodic router
  - Very simple (100s of LoC) A* router with one distinguishing feature: support for per net/per bus periodic boundaries
  - Obeys arbitrary symmetry constraints, allowing us to “think flat” but maintain perfectly tiled design

MDGRAPE-4 Floorplanning

- Prototyped a number of bus layouts using a periodic router
  - Very simple (100s of LoC) A* router with one distinguishing feature: support for per net/per bus periodic boundaries
  - Obeys arbitrary symmetry constraints, allowing us to “think flat” but maintain perfectly tiled design
- Swept many design variables rapidly

MDGRAPE-4 Floorplanning

- Prototyped a number of bus layouts using a periodic router
  - Very simple (100s of LoC) A* router with one distinguishing feature: support for per net/per bus periodic boundaries
  - Obeys arbitrary symmetry constraints, allowing us to “think flat” but maintain perfectly tiled design
- Swept many design variables rapidly
  - Number of NoC layers (2 or 3), router radix (4, 8, 10, 12, 16), bus width (8 ˜ 64 lines), line width (200nm ˜ 600nm)

MDGRAPE-4 Floorplanning

- Prototyped a number of bus layouts using a periodic router
  - Very simple (100s of LoC) A* router with one distinguishing feature: support for per net/per bus periodic boundaries
  - Obeys arbitrary symmetry constraints, allowing us to “think flat” but maintain perfectly tiled design
- Swept many design variables rapidly
  - Number of NoC layers (2 or 3), router radix (4, 8, 10, 12, 16), bus width (8 ˜ 64 lines), line width (200nm ˜ 600nm)
- Picked most promising layouts (minimum maximum length, turn and via counts etc) to investigate further

MDGRAPE-4 Link Planning

- At this stage (2006), we were considering conventional, possibly low-swing differential buses

MDGRAPE-4 Link Planning

- At this stage (2006), we were considering conventional, possibly low-swing differential buses
- Looking at 2.5Gbps ~ 10Gbps signalling:

MDGRAPE-4 Link Planning

- At this stage (2006), we were considering conventional, possibly low-swing differential buses
- Looking at 2.5Gbps ˜ 10Gbps signalling:
  - Firebreathing mm-wave links (à la Dobkin) arbitrarily dismissed as too scary

MDGRAPE-4 Link Planning

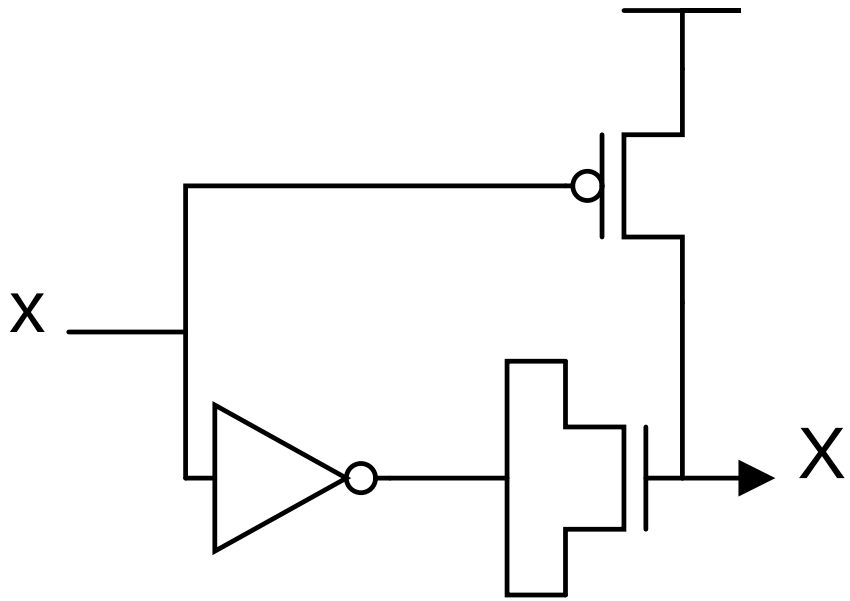
- At this stage (2006), we were considering conventional, possibly low-swing differential buses
- Looking at 2.5Gbps ~ 10Gbps signalling:
  - Firebreathing mm-wave links (à la Dobkin) arbitrarily dismissed as too scary
  - Thicker semi-global/global wiring wasted on slower links

MDGRAPE-4 Link Planning

- At this stage (2006), we were considering conventional, possibly low-swing differential buses
- Looking at 2.5Gbps ~ 10Gbps signalling:
  - Firebreathing mm-wave links (à la Dobkin) arbitrarily dismissed as too scary
  - Thicker semi-global/global wiring wasted on slower links
- Settled on capacitively signalled LVS diff pairs (following Mensink et al.)

Early MDGRAPE-4 Links

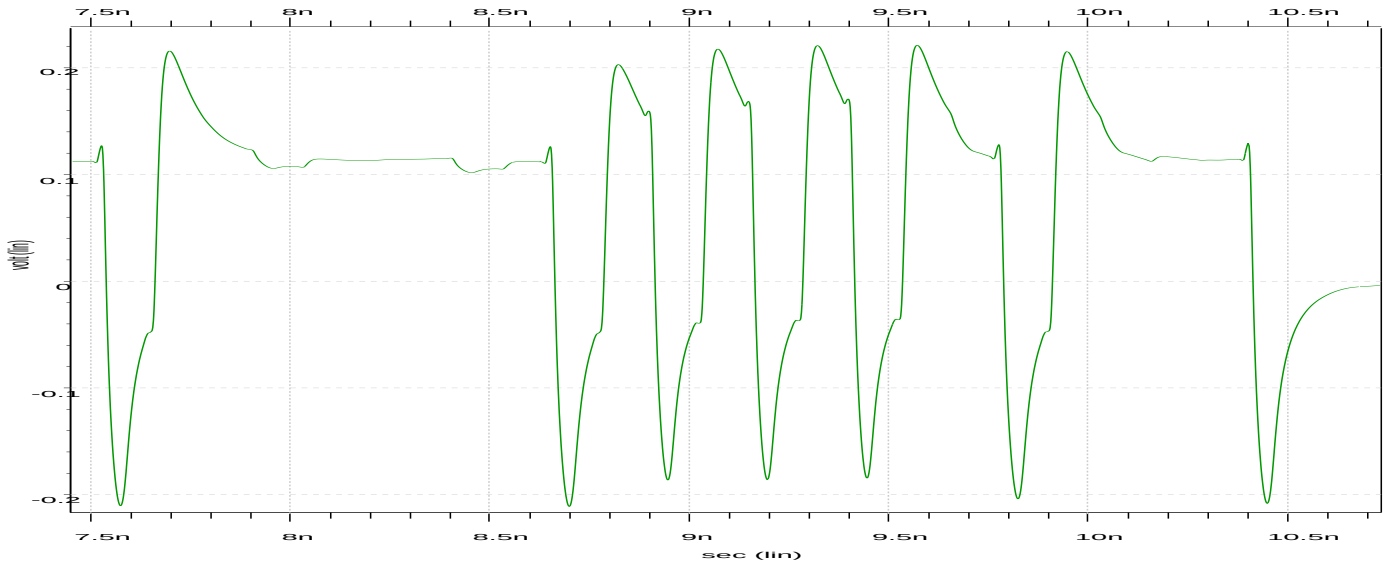
- Basic idea: use capacitive transmitter to save power and boost effective bandwidth



- Set DC level with terminator, but no real DC current flow

Early MDGRAPE-4 Links

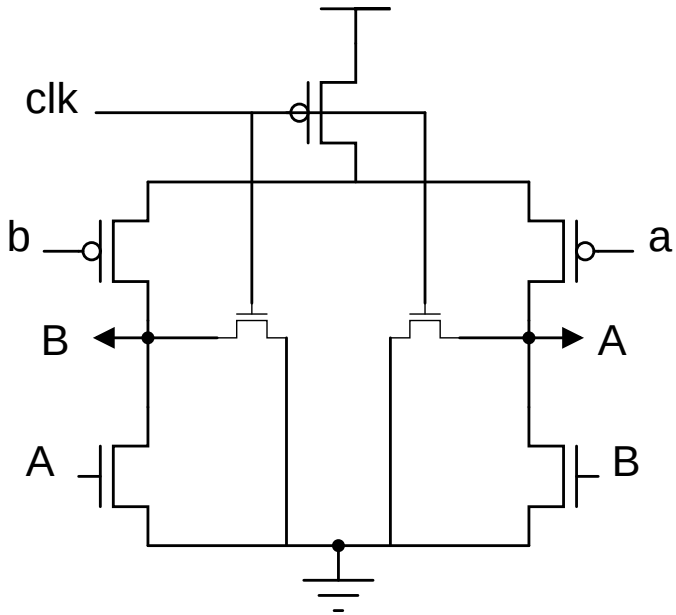
- Basic idea: use capacitive transmitter to save power and boost effective bandwidth



- Set DC level with terminator, but no real DC current flow

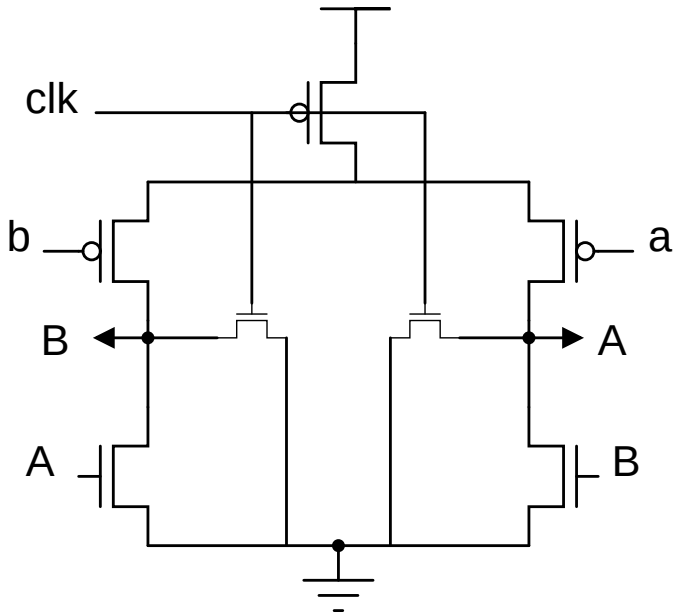
Early MDGRAPE-4 Links

- On the receive side, use sense-amp to restore attenuated signal



Early MDGRAPE-4 Links

- On the receive side, use sense-amp to restore attenuated signal



- "Problem": 4.5mm link performance/power rolled off at ~5GHz

Early MDGRAPE-4 Links

- Investigated a number of different solutions:

Early MDGRAPE-4 Links

- Investigated a number of different solutions:
  - Vary TX/RX power budget

Early MDGRAPE-4 Links

- Investigated a number of different solutions:
  - Vary TX/RX power budget
  - Add repeaters

Early MDGRAPE-4 Links

- Investigated a number of different solutions:
  - Vary TX/RX power budget
  - Add repeaters
  - Place parts of 4.5mm links on lower-k layers

Early MDGRAPE-4 Links

- Investigated a number of different solutions:
  - Vary TX/RX power budget
  - Add repeaters
  - Place parts of 4.5mm links on lower-k layers
  - Active, passive RX-side equalization

Early MDGRAPE-4 Links

- Investigated a number of different solutions:
  - Vary TX/RX power budget
  - Add repeaters
  - Place parts of 4.5mm links on lower-k layers
  - Active, passive RX-side equalization
- All these cost significant power and/or area

MDGRAPE-4 Links

- Feeling pretty sad until one morning in the shower…

MDGRAPE-4 Links

- Feeling pretty sad until one morning in the shower...
- Decided to signal one-hot bit pairs on four-wire bundles:

Data	Encoding
00	0 0 0 1
01	0 0 1 0
10	0 1 0 0
11	1 0 0 0

MDGRAPE-4 Links

- Feeling pretty sad until one morning in the shower…
- Decided to signal one-hot bit pairs on four-wire bundles:

Data	Encoding
00	0 0 0 1
01	0 0 1 0
10	0 1 0 0
11	1 0 0 0

- Called this “quadrature signalling” for want of a better name

MDGRAPE-4 Links

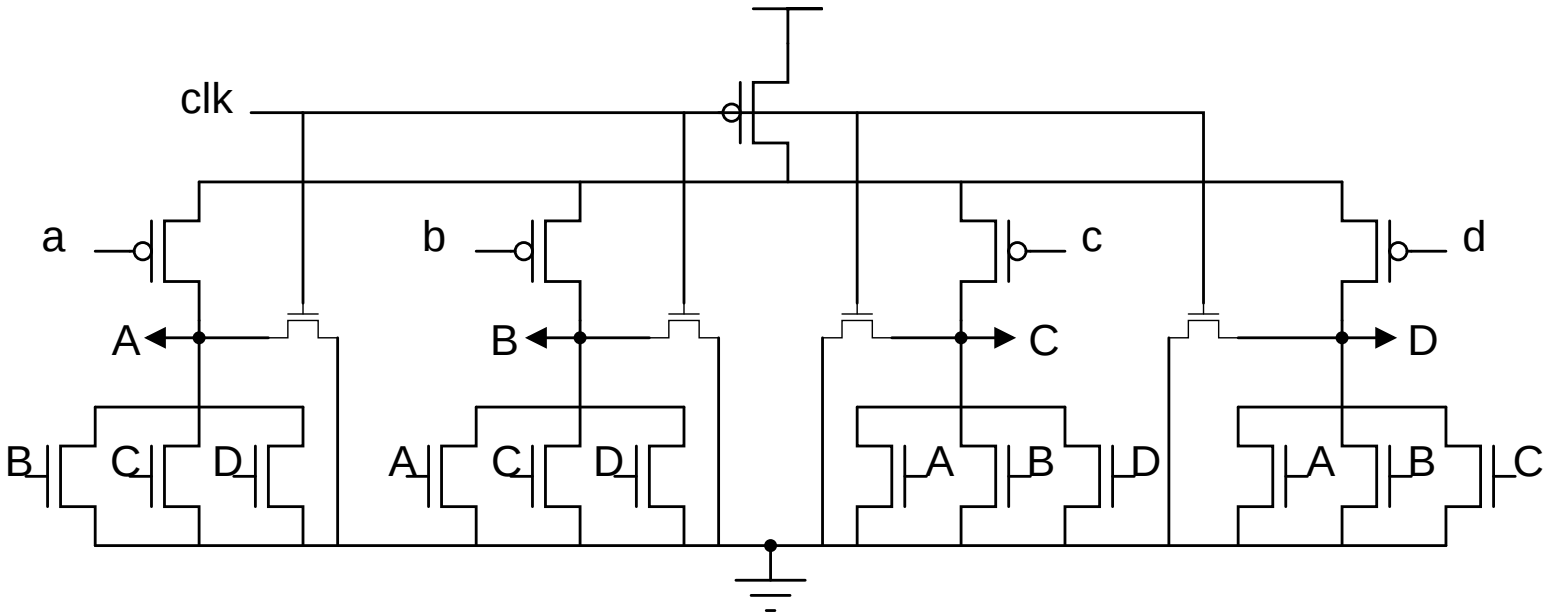
- Feeling pretty sad until one morning in the shower…
- Decided to signal one-hot bit pairs on four-wire bundles:

Data	Encoding
00	0 0 0 1
01	0 0 1 0
10	0 1 0 0
11	1 0 0 0

- Called this “quadrature signalling” for want of a better name
- TX trivial, but how to sense?

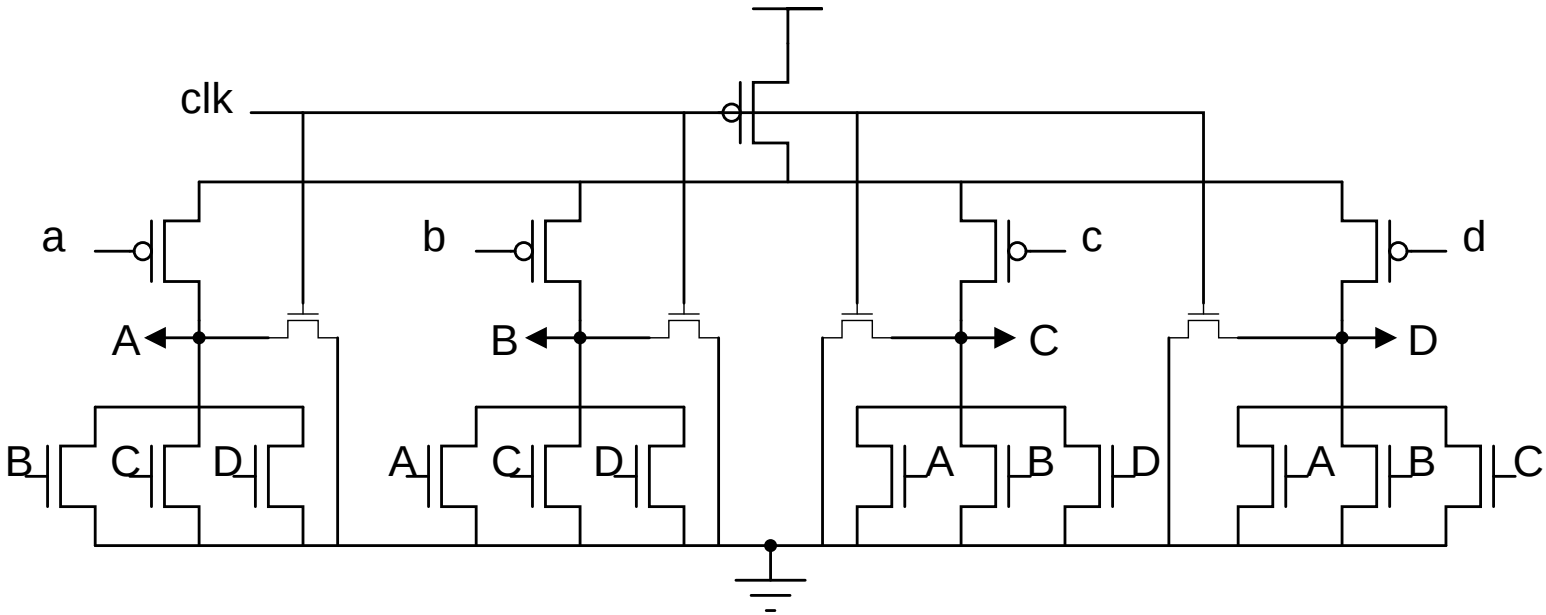
MDGRAPE-4 Links

- It turns out that you can build a sense-amp to sense quadrature signals. With a bit of care, the following circuit works reliably:



MDGRAPE-4 Links

- It turns out that you can build a sense-amp to sense quadrature signals. With a bit of care, the following circuit works reliably:



- Just need to watch out for pitfalls

Pitfall: Variation

- Variation is a real problem

Pitfall: Variation

- Variation is a real problem
  - Four critical FETs instead of two

Pitfall: Variation

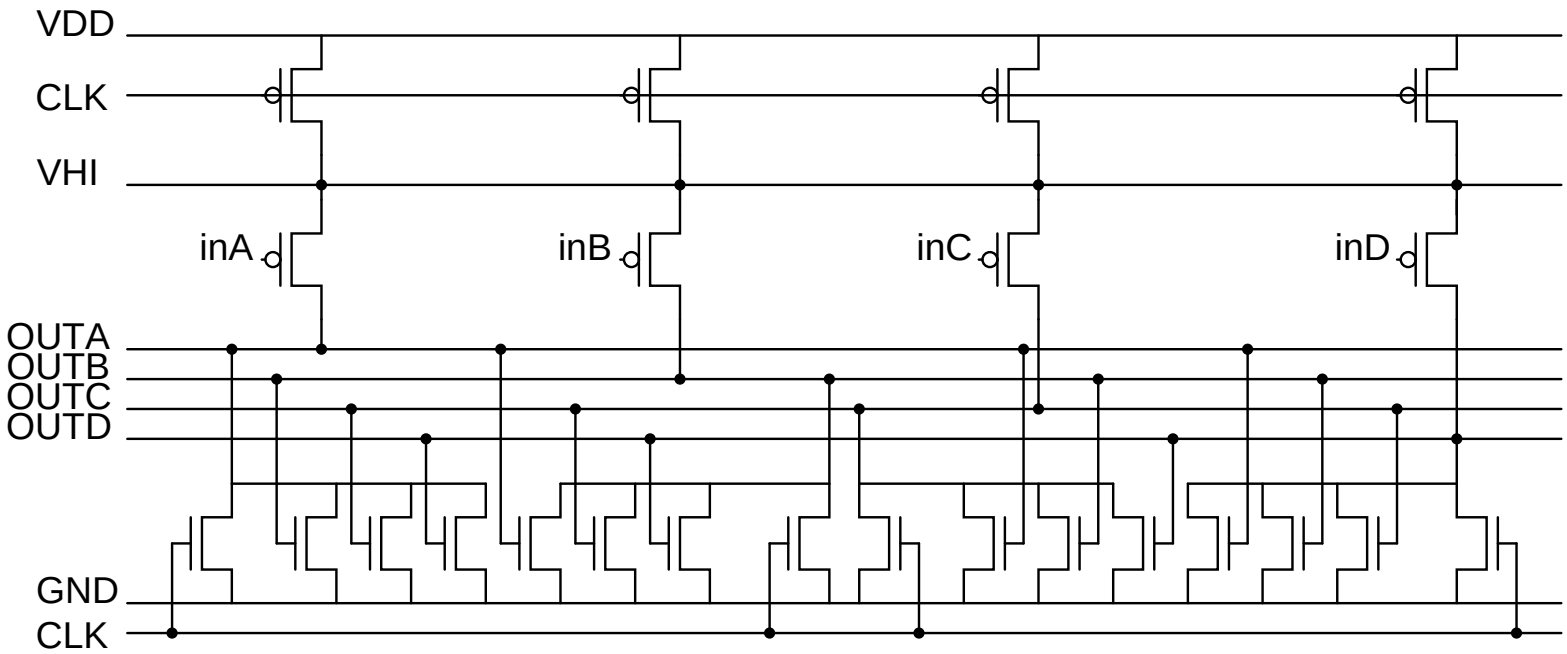
- Variation is a real problem
  - Four critical FETs instead of two
  - Even with those perfectly matched, increased sensitivity to skew forces tighter matching on TX side also

Pitfall: Variation

- Variation is a real problem
  - Four critical FETs instead of two
  - Even with those perfectly matched, increased sensitivity to skew forces tighter matching on TX side also
- Layout critical on both ends

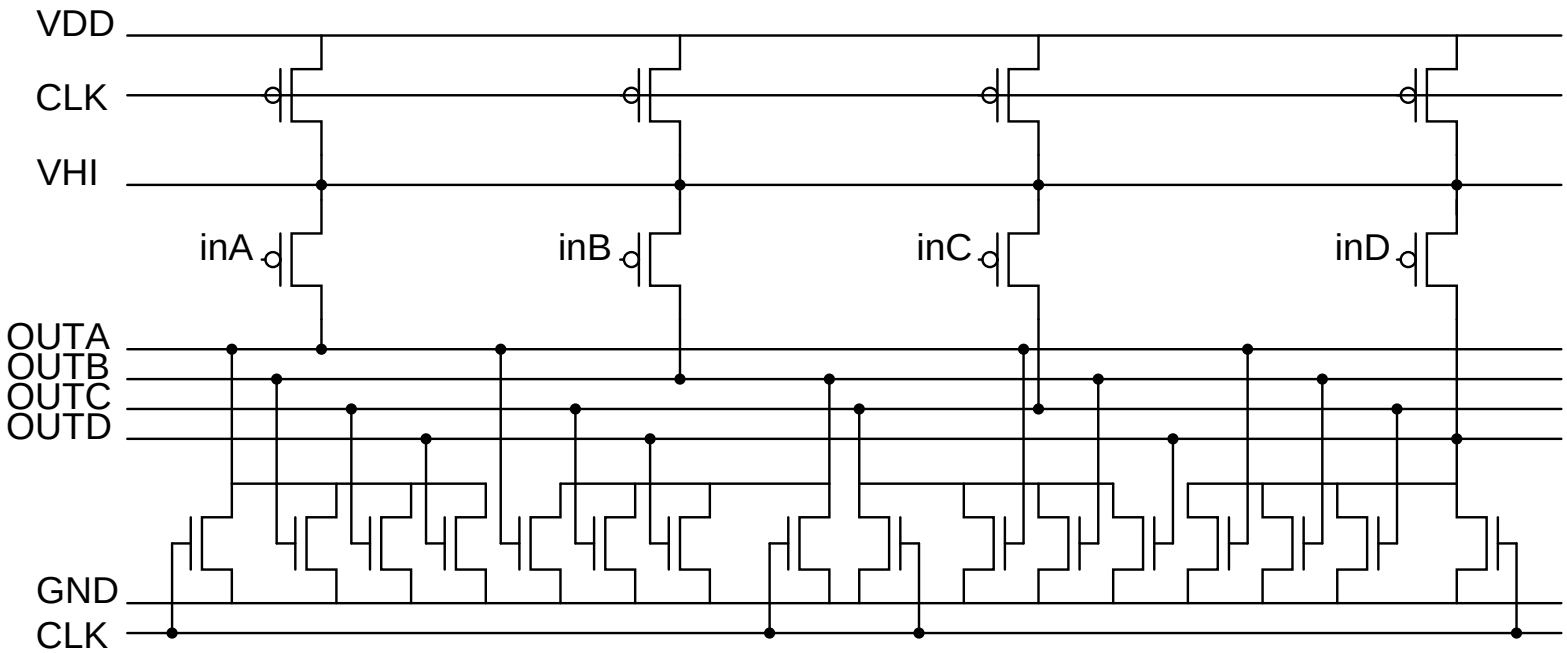
Pitfall: Layout

- QSA layout is more or less as in this schematic:



Pitfall: Layout

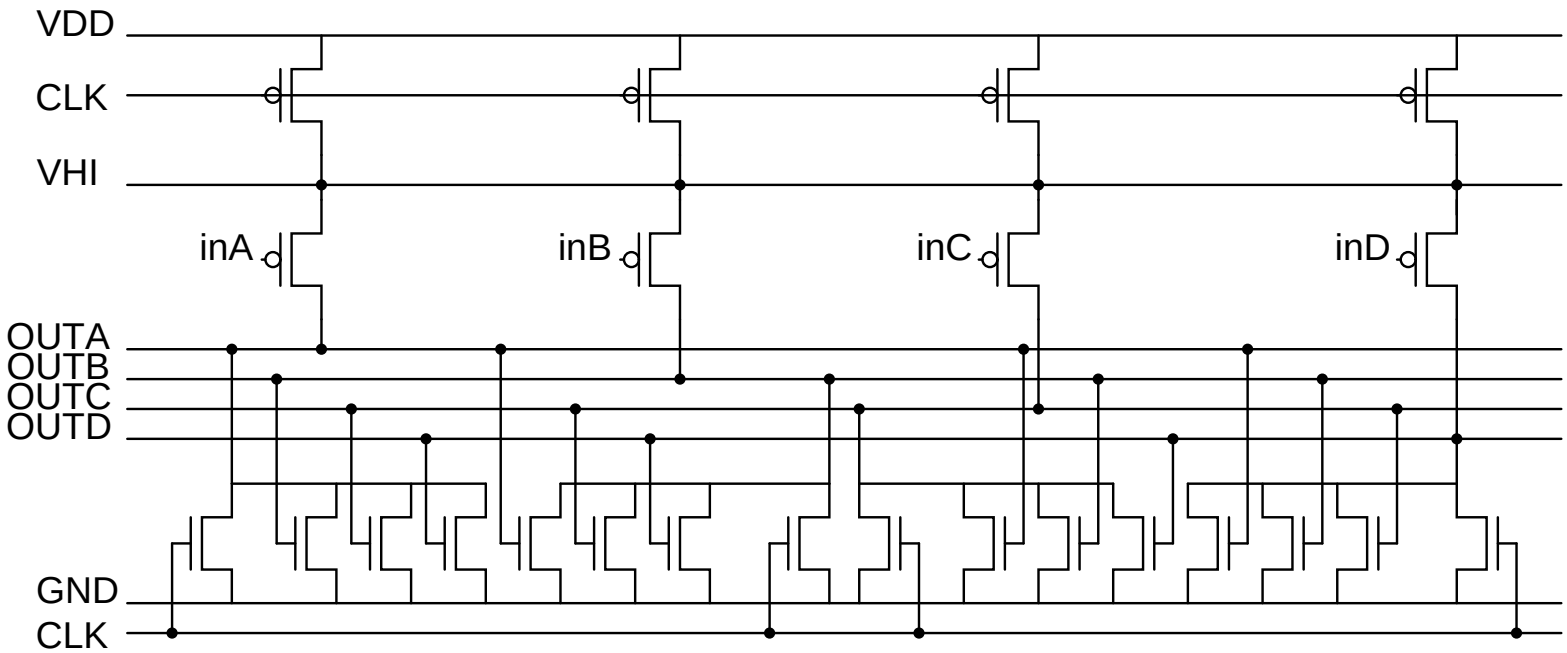
- QSA layout is more or less as in this schematic:



- Uniform, n-wavelength poly pitches improve matching

Pitfall: Layout

- QSA layout is more or less as in this schematic:



- Uniform, n-wavelength poly pitches improve matching
- TX similar: simple arrays of identical devices

Pitfall: Vcm

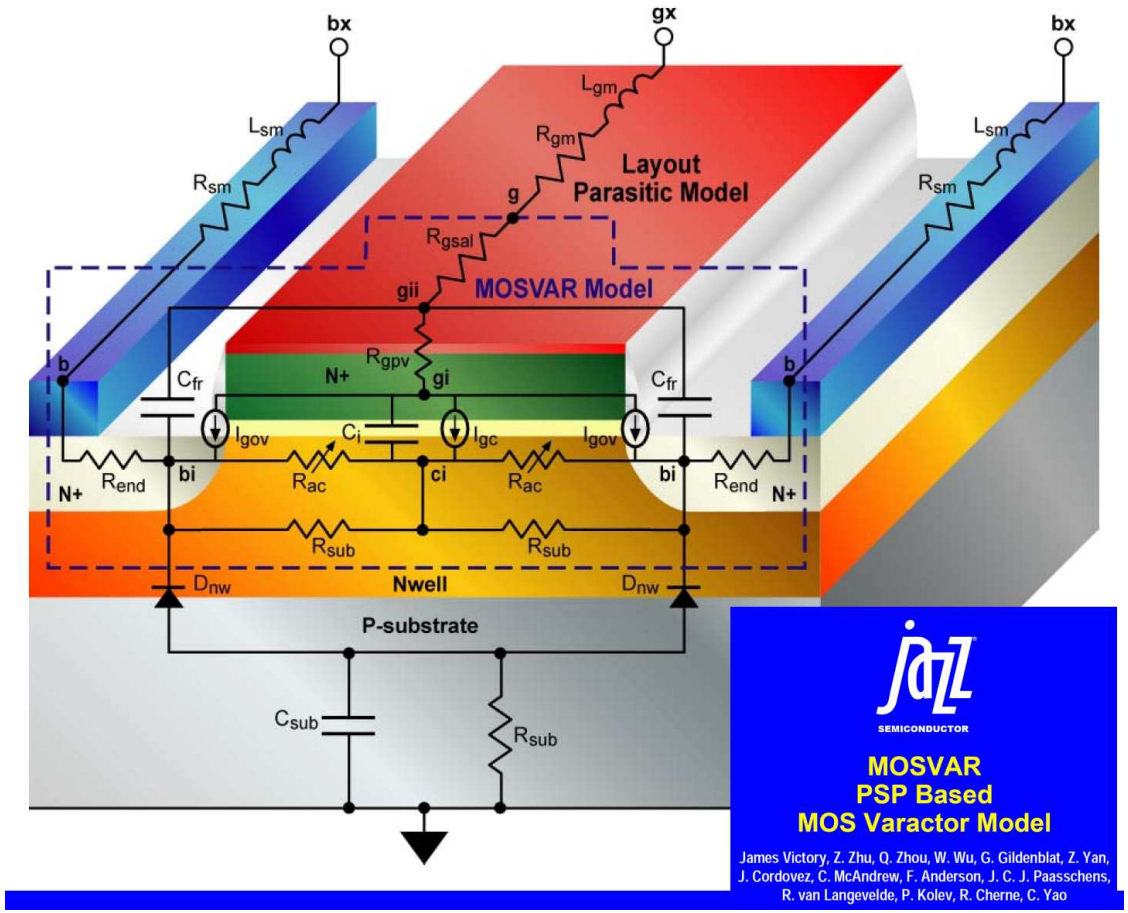
- A key variation-related decision: signal around vdd or gnd?

Pitfall: Vcm

- A key variation-related decision: signal around vdd or gnd?
- Each has pros and cons...

Pitfall: Vcm

- Signalling around vdd gets higher capacitance and lower variation from the TX varactors:



Pitfall: Vcm

- However, signalling around gnd enables the use of PMOS sense FETs in the receiver, allowing significantly lower offsets.

Pitfall: Vcm

- However, signalling around gnd enables the use of PMOS sense FETs in the receiver, allowing significantly lower offsets.
- In our situation, this turned out to be the more important effect.

Pitfall: Other Variation Effects

- In addition to getting RX and TX circuits right, need to keep link skew down. Two main effects identified:

Pitfall: Other Variation Effects

- In addition to getting RX and TX circuits right, need to keep link skew down. Two main effects identified:
  - Intrinsic layout: required development of a simple tuner to adjust twist locations and add extra length (serpentine) to match wire loads within and between buses

Pitfall: Other Variation Effects

- In addition to getting RX and TX circuits right, need to keep link skew down. Two main effects identified:
  - Intrinsic layout: required development of a simple tuner to adjust twist locations and add extra length (serpentine) to match wire loads within and between buses
  - Thermal: planning for 30 °C temperature deltas leads to nearly 10% resistance change

Pitfall: Other Variation Effects

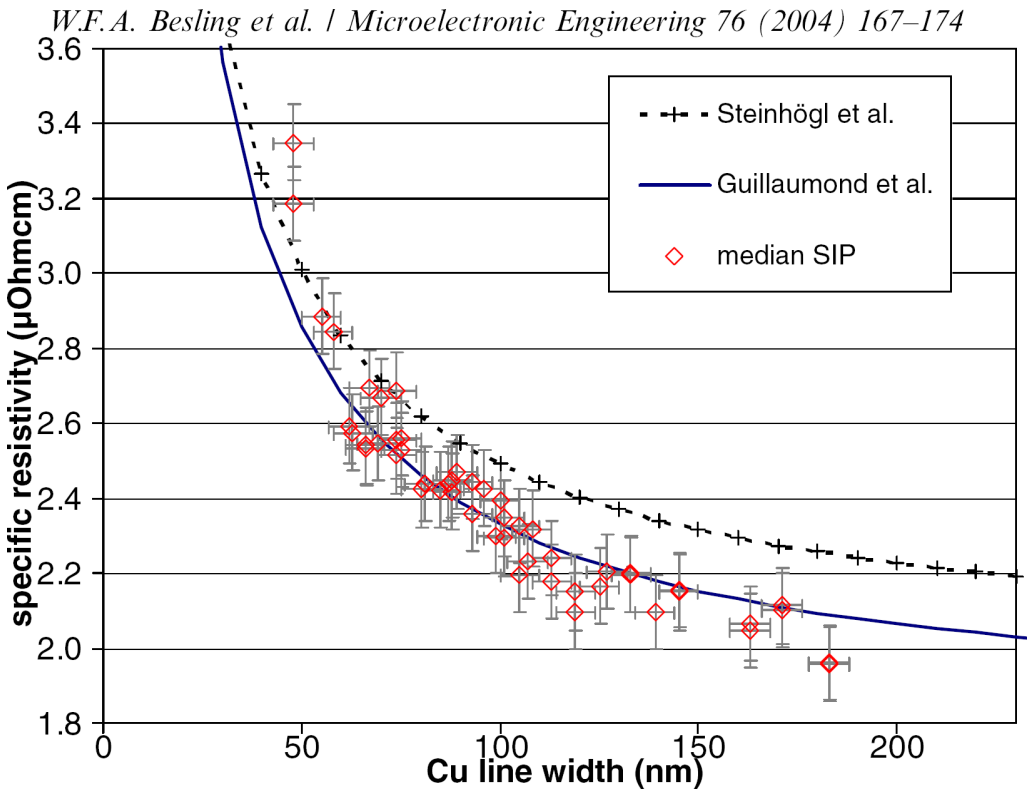
- In addition to getting RX and TX circuits right, need to keep link skew down. Two main effects identified:
  - Intrinsic layout: required development of a simple tuner to adjust twist locations and add extra length (serpentine) to match wire loads within and between buses
  - Thermal: planning for 30 °C temperature deltas leads to nearly 10% resistance change
    - Smaller effect than on TX/RX blocks, but too large to ignore

Pitfall: Other Variation Effects

- In addition to getting RX and TX circuits right, need to keep link skew down. Two main effects identified:
  - Intrinsic layout: required development of a simple tuner to adjust twist locations and add extra length (serpentine) to match wire loads within and between buses
  - Thermal: planning for 30 °C temperature deltas leads to nearly 10% resistance change
    - Smaller effect than on TX/RX blocks, but too large to ignore
- Thermal and cross-wafer/lot-to-lot variation handled through per-bus clock verniers

Pitfall: Other Wire Effects

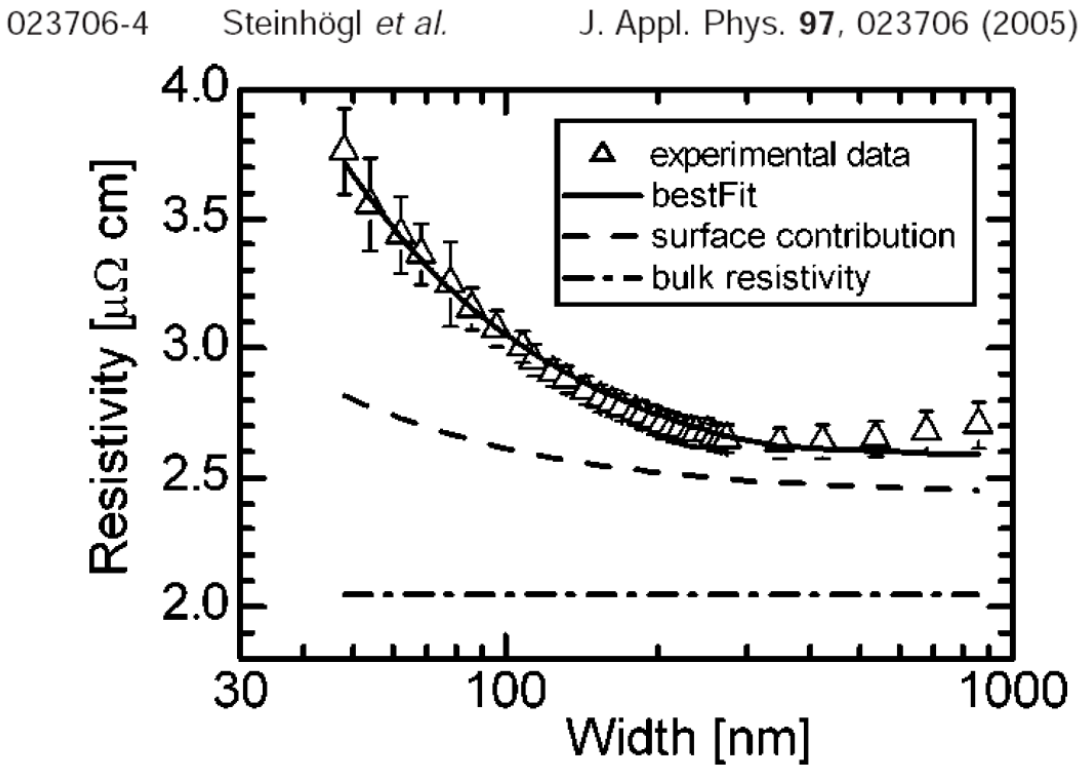
- Usually suppose that more, thinner links will yield a better result
- Often true today, but problems with wire scaling motivate fatter serial links:



- 350nm links give close-to-bulk resistivity and low variation

Pitfall: Other Wire Effects

- Usually suppose that more, thinner links will yield a better result
- Often true today, but problems with wire scaling motivate fatter serial links:



- 350nm links give close-to-bulk resistivity and low variation

Pitfall: Yield

- Despite best efforts in TX and RX cell design, yield still low

Pitfall: Yield

- Despite best efforts in TX and RX cell design, yield still low
- A function of bit error rate, difficult and/or timeconsuming to estimate

Pitfall: Yield

- Despite best efforts in TX and RX cell design, yield still low
- A function of bit error rate, difficult and/or timeconsuming to estimate
  - But simple, highly regular layout diminishes a number of layout-dependent effects (well proximity, stress, OPC) leaving basic $(WL)^{1/2}$-like random variations more dominant

Pitfall: Yield

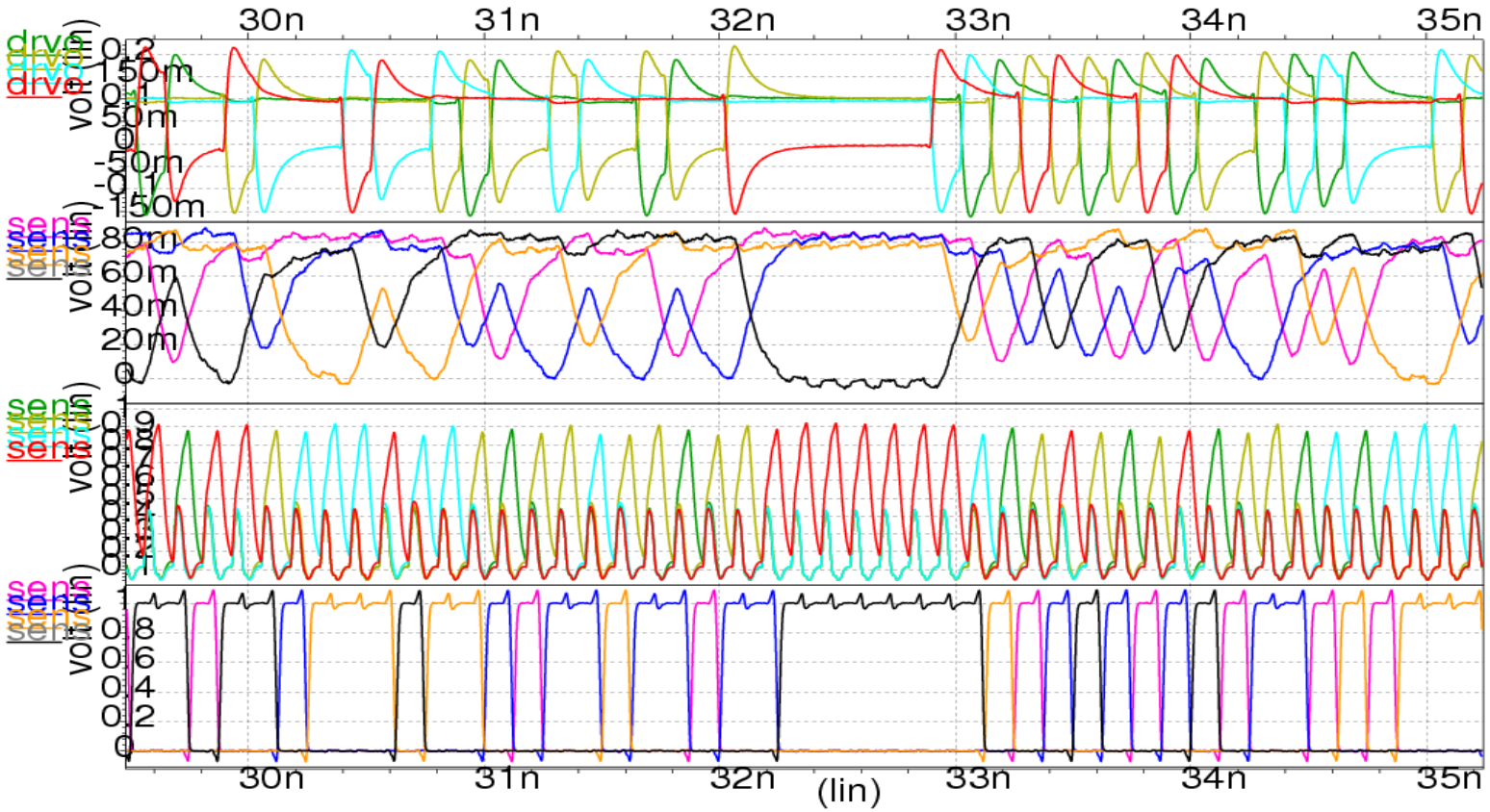
- Despite best efforts in TX and RX cell design, yield still low
- A function of bit error rate, difficult and/or timeconsuming to estimate
  - But simple, highly regular layout diminishes a number of layout-dependent effects (well proximity, stress, OPC) leaving basic $(WL)^{1/2}$-like random variations more dominant
- Essential to insert spare bus lines

Pitfall: Yield

- Despite best efforts in TX and RX cell design, yield still low
- A function of bit error rate, difficult and/or timeconsuming to estimate
  - But simple, highly regular layout diminishes a number of layout-dependent effects (well proximity, stress, OPC) leaving basic $(WL)^{1/2}$-like random variations more dominant
- Essential to insert spare bus lines
  - QSAs small and reliable enough (temporally speaking) that adding spares is more effective than adding sophisticated ECC from a power/performance standpoint

Result

- Result: Quadrature TX/RX enables successful 8GHz operation across 4.5mm links
  - With low area.. (TX essentially unchanged, RX reduced due to elimination of EQ)
  - ..and low power (61fJ/bit typical at 8GHz in 1.0V 65nm)



Future Work

- A number of variants on the sense-amp are possible (as usual)

Future Work

- A number of variants on the sense-amp are possible (as usual)
  - Static (as opposed to clocked) load device costs power, but greatly reduces kickback and parasitic clock noise, and saves some clock power (will be more useful with short-channel 45nm, 32nm)

Future Work

- A number of variants on the sense-amp are possible (as usual)
  - Static (as opposed to clocked) load device costs power, but greatly reduces kickback and parasitic clock noise, and saves some clock power (will be more useful with short-channel 45nm, 32nm)
  - Can push this even further (peaking loads, body biasing)

Future Work

- A number of variants on the sense-amp are possible (as usual)
  - Static (as opposed to clocked) load device costs power, but greatly reduces kickback and parasitic clock noise, and saves some clock power (will be more useful with short-channel 45nm, 32nm)
  - Can push this even further (peaking loads, body biasing)
- Lower-speed, zero static power variants will be more efficient than standard differential circuits in highly power concious/low voltage situations

Future Work

- A number of variants on the sense-amp are possible (as usual)
  - Static (as opposed to clocked) load device costs power, but greatly reduces kickback and parasitic clock noise, and saves some clock power (will be more useful with short-channel 45nm, 32nm)
  - Can push this even further (peaking loads, body biasing)
- Lower-speed, zero static power variants will be more efficient than standard differential circuits in highly power concious/low voltage situations
  - Plenty of low-power quadrature sense-amp topologies to explore...

Future Work

- A number of variants on the sense-amp are possible (as usual)
  - Static (as opposed to clocked) load device costs power, but greatly reduces kickback and parasitic clock noise, and saves some clock power (will be more useful with short-channel 45nm, 32nm)
  - Can push this even further (peaking loads, body biasing)
- Lower-speed, zero static power variants will be more efficient than standard differential circuits in highly power concious/low voltage situations
  - Plenty of low-power quadrature sense-amp topologies to explore...
- A number of well-known equalizer designs plug directly into a QSA: where yield and/or area are much less important than absolute speed, these deliver significant gains, sometimes at modest power costs

Future Work

- Periodic router turned out to be more useful than we expected

Future Work

- Periodic router turned out to be more useful than we expected
  - Saved a lot of time during early design space exploration

Future Work

- Periodic router turned out to be more useful than we expected
  - Saved a lot of time during early design space exploration
  - Kept saving minutes here and there by shrinking GDS files, detail router and extraction runtimes

Future Work

- Periodic router turned out to be more useful than we expected
  - Saved a lot of time during early design space exploration
  - Kept saving minutes here and there by shrinking GDS files, detail router and extraction runtimes
- A design methodology gaining traction recently is “let’s place the cells uniformly and then just let the router do its thing” (witness new relative placement commands in IC Compiler, Matlab cell layout scripts, etc)

Future Work

- Periodic router turned out to be more useful than we expected
  - Saved a lot of time during early design space exploration
  - Kept saving minutes here and there by shrinking GDS files, detail router and extraction runtimes
- A design methodology gaining traction recently is “let’s place the cells uniformly and then just let the router do its thing” (witness new relative placement commands in IC Compiler, Matlab cell layout scripts, etc)
- This works fine, but we can potentially improve router and extraction runtimes if placement/symmetry information can be communicated to these tools

Summary

- MDGRAPE-4 designed from the ground up as a heterogenous, tiled processor

Summary

- MDGRAPE-4 designed from the ground up as a heterogenous, tiled processor
- Early design centered on tightly coupled floorplanning and link length estimation

Summary

- MDGRAPE-4 designed from the ground up as a heterogenous, tiled processor
- Early design centered on tightly coupled floorplanning and link length estimation
- NoC design exploration used periodic wire router to quickly prototype various designs

Summary

- MDGRAPE-4 designed from the ground up as a heterogenous, tiled processor
- Early design centered on tightly coupled floorplanning and link length estimation
- NoC design exploration used periodic wire router to quickly prototype various designs
- One-hot quadrature links save power, reduce area and improve link speed compared to standard differential designs

Acknowledgements

- We'd like to thank Fujitsu Microelectronics, Berkeley Design Automation and Magma Design Automation for their support
- Thanks to Miki Tanaka (AIST) for work on the periodic router
- Thanks also to James Victory (Jazz Semiconductor) for MOSVAR model insights
- We're grateful for funding from Japan's JSPS and RIKEN

Thanks for listening!

Thanks for listening - any questions?

- More bits and pieces online: `http://mdgrape.gsc.riken.jp`