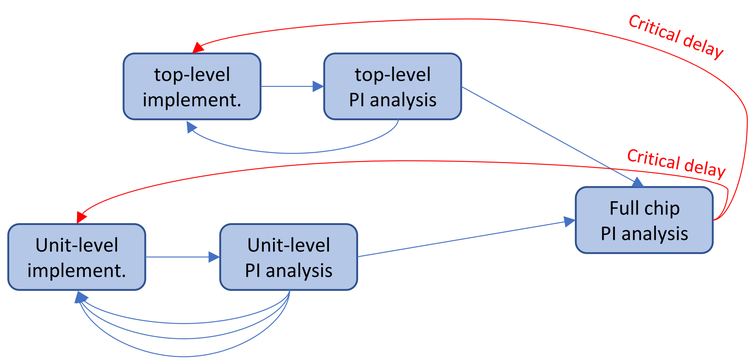
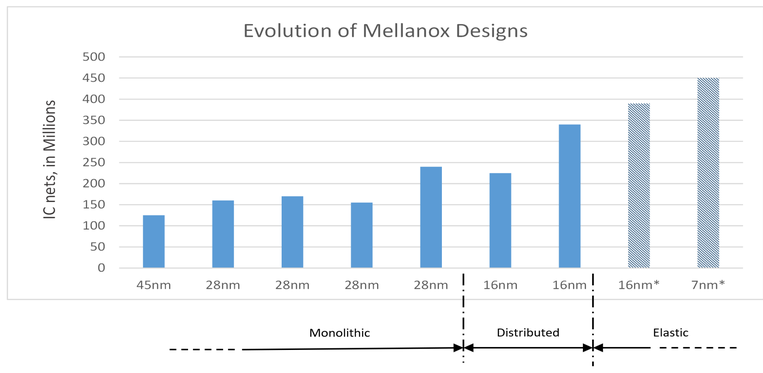
Fast and Accurate Incremental Power and Signal Integrity Analysis

Rotem Cohen, Anton Rozen (Mellanox Technologies)
Abhijith M V, Ramesh Agarwal, Sankar Ramachandran, Scott Johnson (Ansys Inc.)

Motivation

- Grid complexity and number of gates are increasing dramatically per technology over the years.
- A fast turn around time with razor thin voltage accuracy is necessary to ensure voltage and timing signoff.
- The power analysis tools need to evolve accordingly, in order to face IC design complexity - starting from stand-alone engine, through multi-machine solution, towards to distributed calculations and "big-data" management.
- Considerable effort is spend on unit-level analysis runs and top-level analysis runs (hours and days long, respectively). ECO mode usually requires full-chip simulations that are both fast and high-accuracy.



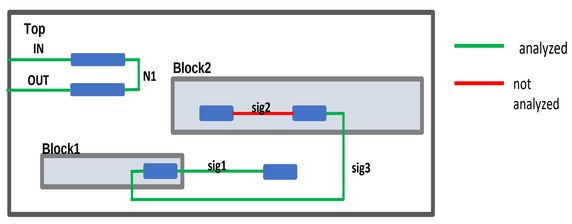


Architecture	Generation	Number of machines	Max Node Count that can be handled	Example
Monolithic	First – 2003 onwards	1	1B	RedHawk™ & others
Distributed	Second – 2013/14	32	4B	RedHawk-DMP™ & others
Elastic	Third – 2016/18 onwards	Scalable beyond 1000 cores also	Unlimited and Scalable	RedHawk-SC™

Incremental analysis (power abstraction)

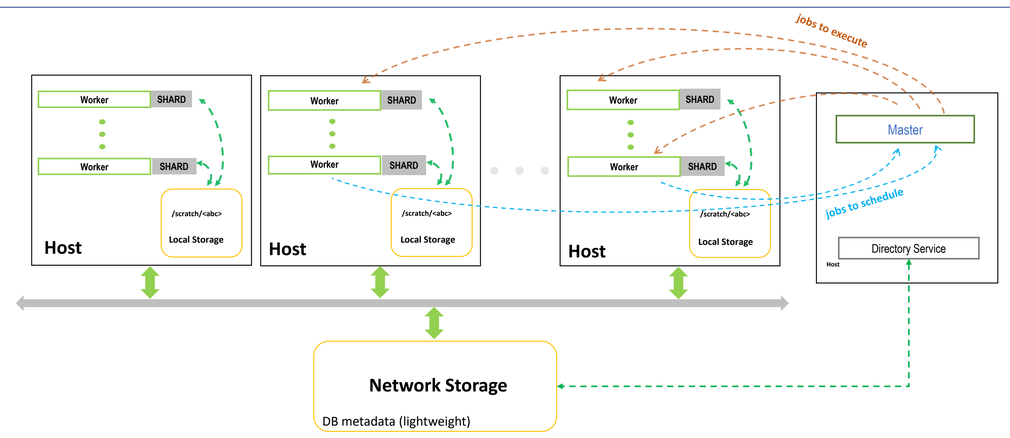
- **EM/IR Analysis - Bottom-top power grid abstraction**
- In order to speed up full-chip IR drop simulations, Power Integrity simulation may use abstraction (roll-up) of low- and mid-level metals power grid (M1…M11).
- Such abstraction can be used in full-chip simulations, and design should fulfill several requirements.
- Abstraction of Power Grid metal layers preferred to be performed on non-shared stripes on unit-independent power grids. Shared top-level metals should stay in flat views.
- Power distribution of abstracted area / unit should be uniform.

- **Signal EM – top level signals simulation**
- When signals view reduction is performed, it's highly important to fulfil two contradictive requirements:
  1. Remove all in-unit signals in order to reduce analysis complexity
  2. Preserve unit's boundary signals in order to analyze correctly paths of "unit=>top-level" and "unit=>full chip=>unit"



Net	Top only	flat
Top level net (N1)	✓	✓
Primary input net (IN)	✓	✓
Primary output net (OUT)	✓	✓
Interface net from block1 to Top (sig1)	✓	✓
Interface net from block2 to block1 through Top (sig3)	✓	✓
Block level net inside block2 (sig2)	×	✓

Distributed computing

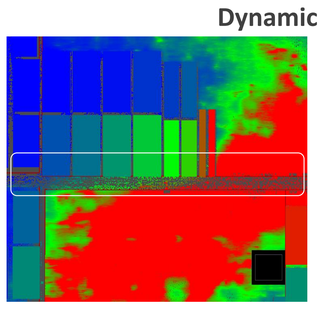


- Big data analytics enable rapid data mining and analytics to drive actionable outcomes and optimization.
- The **Elastic Computing** ability helps in processing each scenario in parallel, or in series, depending on amount of CPU cores available at that time. The amount of cores can be dynamically adjusted, based on need and availability.
- **Workers** are job execution daemons (processes) that communicate to others/master using TCP/IP sockets.
- **Shard** is a horizontal partition of data in a database or search engine. Each of multiple shards is held on a separate database server instance, to spread load.
- **Complexity**
  - Very large number of hosts (> 1000 workers)
  - All hosts need to communicate with each other
- **Need for local Disk**
  - Each shard is stored directly on local disk of the Worker
  - R/W access to local storage has no network impact

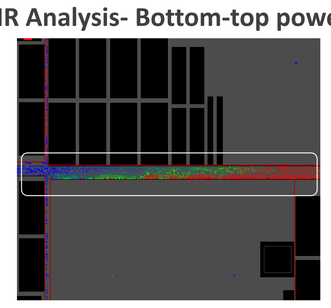
Results & Conclusions

Incremental Analysis Comparison with Flat Analysis

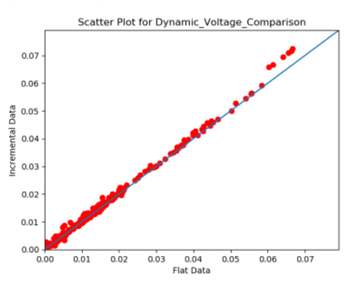
Dynamic IR Analysis- Bottom-top power abstraction



Top-level flat Run

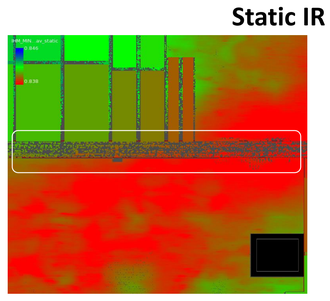


Top-level incremental Run

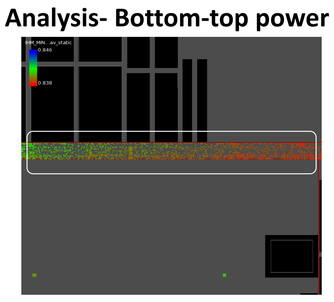


Dynamic Voltage Drop Scatter Plot

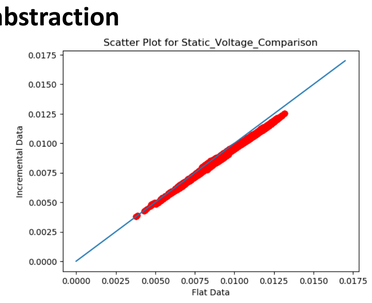
Static IR Analysis- Bottom-top power abstraction



Top-level flat Run

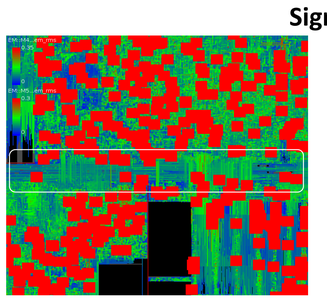


Top-level incremental Run

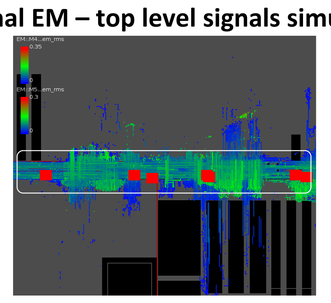


Static Voltage Drop Scatter Plot

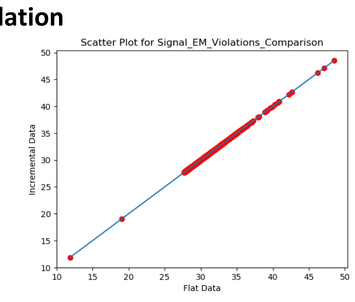
Signal EM – top level signals simulation



Top-level flat Run



Top-level incremental Run



Signal EM Scatter Plot

Scalability Comparison

	Monolithic	Distributed	Full-chip Elastic Computing
Technology	28nm	16nm	16nm
Chip size	1/4th Fullchip 96M nets	Fullchip 225M nets	Fullchip 340M nets
CPU core usage / Machines	1 machine of 1TB	4 machines of 1.4TB	150 works of 72GB
Run time	60Hrs	72Hrs	24Hrs

	Item	Flat Analysis	Incremental Analysis (using abstraction)
cluster-level	# CPU	23	7
	Total Memory Usage	23x25GB	7x25GB
	Disc space	766GB	341GB
	Run time	10.5Hrs	5.5Hrs
Full-chip	# CPU	447	152
	Total Memory Usage	447x25GB**	152x25GB
	Disc space	~13.5TB**	6TB
	Run time	~50Hrs**	26Hrs

**extrapolation

Conclusions

- Using elastic computing and power abstraction, resulted in **3x improvement** in runtime of static / dynamic / sigEm analysis.
- Incremental power gave reasonable correlation with the flat analysis and the results were correlating:
  - **Static 4%,**
  - **Dynamic 2.5%**
  - **Signal EM perfectly correlating**.
- Top level analysis iterations gain speedup by avoiding reiterating unit-level runs.

What's next?

Limitations

- The selection of metal layer to which the abstraction to be done is manual and user needs the knowledge of the design.
- The incremental PG analysis is less effective for wire bond designs (large ASIC designs are usually not wirebond, but flip-chip).
- The bottom up power abstraction is layout based, so unconnected instances might still receive fault sinks at top of abstraction.
- If adjacent blocks have power grid, connected by lower (abstracted) metal layer/s, accuracy of the abstraction method may degrade.

Future Scopes

- Automatic layer selection can be done as a future enhancement.
- Abstraction sinks connections will be circuit-based rather than layout-based (preventing fault connections).