Big Data Analytics
An Introduction

Giuseppe Nicosia
Dept. of Mathematics & Computer Science, University of Catania, Italy
www.dmi.unict.it./nicosia/
nicosia@dmi.unict.it

What is “Big”? And What is Data?

What is “Big”? Peta, exa, zetta, yotta

Specific units of IEC 60027-2 A.2 and ISO/IEC 80000

IEC prefix		Representations				Customary prefix	
Name	Symbol	Base 2	Base 1024	Value Base 10		Name	Symbol
kibi	Ki	2^{10}	1024^{1}	1024	≈ 1.02×10^{3}	kilo	k or K
mebi	Mi	2^{20}	1024^{2}	1048576	≈ 1.05×10^{6}	mega	M
gibi	Gi	2^{30}	1024^{3}	1073741824	≈ 1.07×10^{9}	giga	G
tebi	Ti	2^{40}	1024^{4}	1099511627776	≈ 1.10×10^{12}	tera	T
pebi	Pi	2^{50}	1024^{5}	1125899906842624	≈ 1.13×10^{15}	peta	P
exbi	Ei	2^{60}	1024^{6}	1152921504606846976	≈ 1.15×10^{18}	exa	E
zebi	Zi	2^{70}	1024^{7}	1180591620717411303424	≈ 1.18×10^{21}	zetta	Z
yobi	Yi	2^{80}	1024^{8}	1208925819614629174706176	≈ 1.21×10^{24}	yotta	Y

Data

Data: *1640s*, classical plural of *datum*, from Latin *datum* "(thing) given," neuter past participle of *dare* "**to give**"*.

Data: "**transmittable and storable computer information**" first recorded 1946*.

Data processing is from 1954*.

Data **is a set of values of qualitative or quantitative variables****.

Raw Data, Experimental Data, Digital Data, Analog Data, Open Data, etc.

* http://www.etymonline.com/index.php?term=data

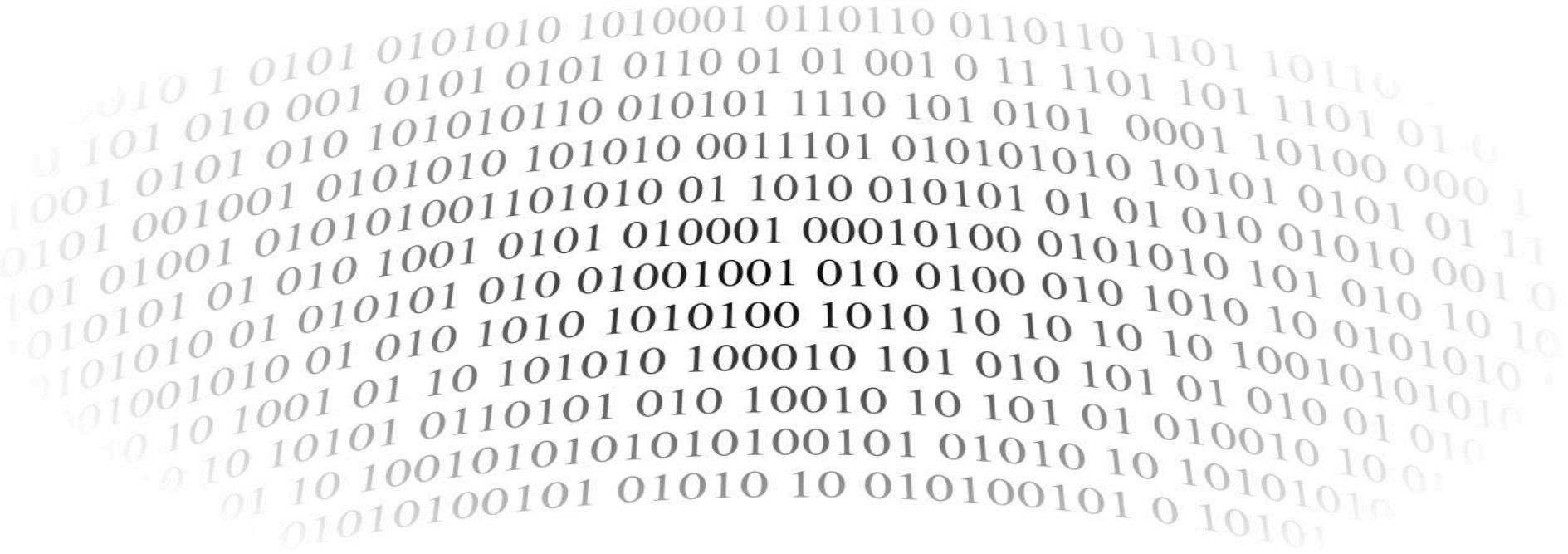
** https://en.wikipedia.org/wiki/Data

Data in Computer Science

Data is any sequence of one or more symbols (from a given set of alphabets) given meaning by specific act(s) of interpretation.

Data is not information.

Data requires interpretation to become information. To translate data to information, there must be several known factors considered. The factors involved are determined by the creator of the data and the desired information.



Definition of Big Data

There is no agreed upon definition for "big data"

3Vs Definition of Big Data

"Big data is **high-volume**, **high-velocity** and **high-variety** information assets that demand cost-effective, innovative forms of information processing forenhanced insight and decision making." *Gartner*

which was derived from:

"While enterprises struggle to consolidate systems and collapse redundant databases to enable greater operational, analytical, and collaborative consistencies, changing economic conditions have made this job more difficult. E-commerce, in particular, has exploded data management challenges along **three dimensions: volumes, velocity and variety**. In 2001/02, IT organizations much compile a variety of approaches to have at their disposal for dealingeach." *Doug Laney*

4Vs Definition of Big Data

- The four dimensions (V's) of Big Data:
  - Volume,
  - Velocity,
  - Variety &
  - **Veracity**.
- Big data is not just about size. Finds insights from complex, noisy, heterogeneous, longitudinal, and voluminous data.
- It aims to answer questions that were previously unanswered.
- The challenges include
  - capturing,
  - Storing,
  - searching,
  - sharing &
  - analyzing.

Another Definition of Big Data

Big data is **data so large that it does not fit in the main memory of a single machine**, and the need to process big data by efficient algorithms arises in Internet search, network traffic monitoring, machine learning, scientific computing, signal processing, and several other areas.

CS 229r: Algorithms for Big Data – Prof. Jelani Nelson, Harvard University

http://people.seas.harvard.edu/~minilek/cs229r/fall15/index.html

Another (?!?) Definition (?!?) of Big Data

- The tools of **data science** are as appropriate for gigabyte (10^9) as they are for petabyte (10^{15}) scale datasets.
- "**Big data**" typically refers to data on the scale of terabytes (10^{12}) and petabytes (10^{15}, a million gigabytes).

https://datascience.berkeley.edu/about/what-is-data-science/

Why Data Scientists?

50x in 2020 the world will generate 50 times the amount of data than in 2011

Source: emc.com

"**This hot new field promises to revolutionize industries from business to government, health care to academia**." *The New York Times*

The statistics listed below represent **this significant and growing demand for data scientists**.

#16 Highest Paying Job in Demand

3,433 Number of Job Openings

$105,395 Average Base Salary

#1 Best Job in America for 2016

Sources: **25 Best Jobs in America**: http://www.glassdoor.com/blog/jobs-america/

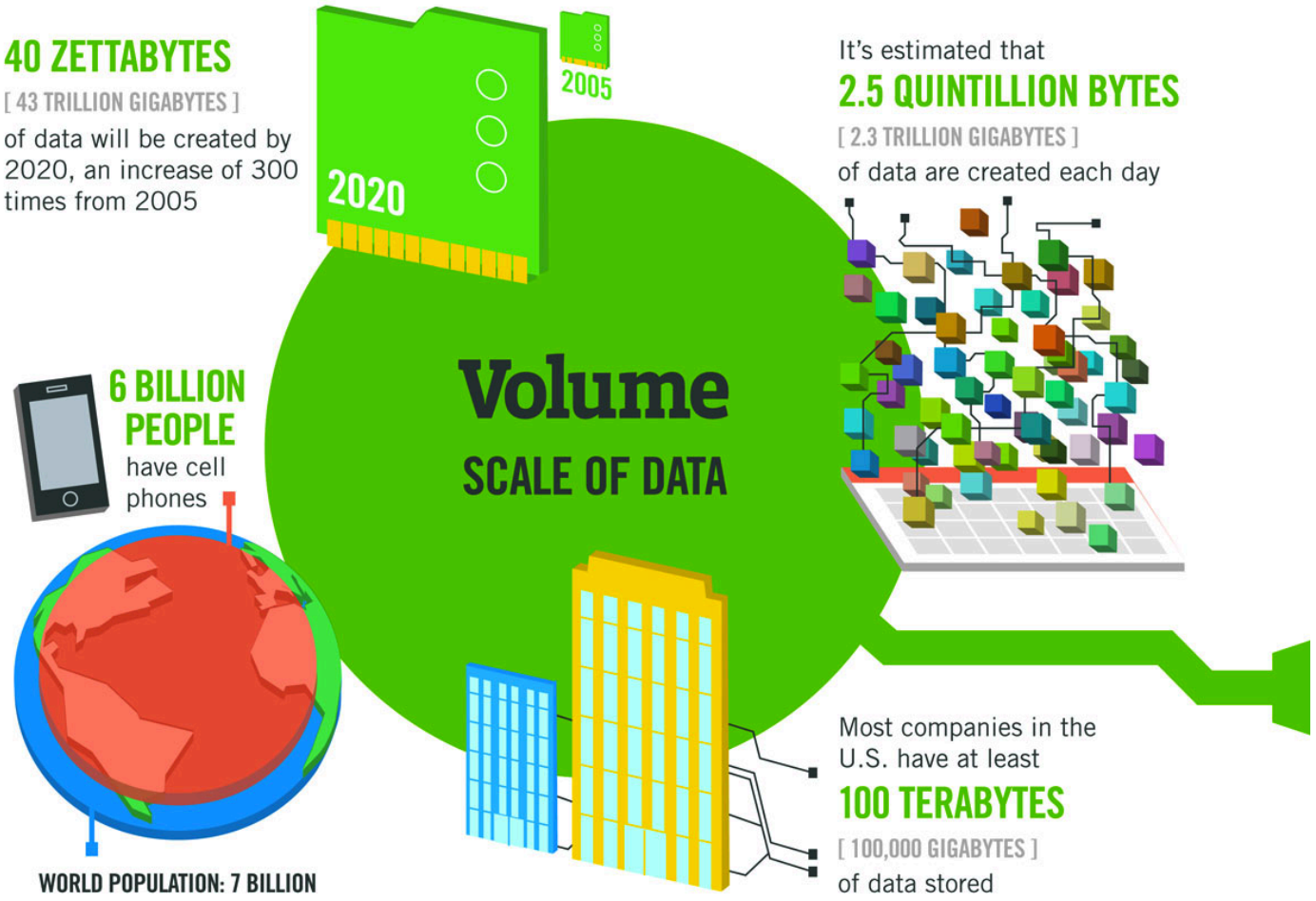
25 Highest Paying Jobs in America for 2016 https://www.glassdoor.com/blog/25-highest-paying-jobs-america-2016/

https://datascience.berkeley.edu/about/what-is-data-science/

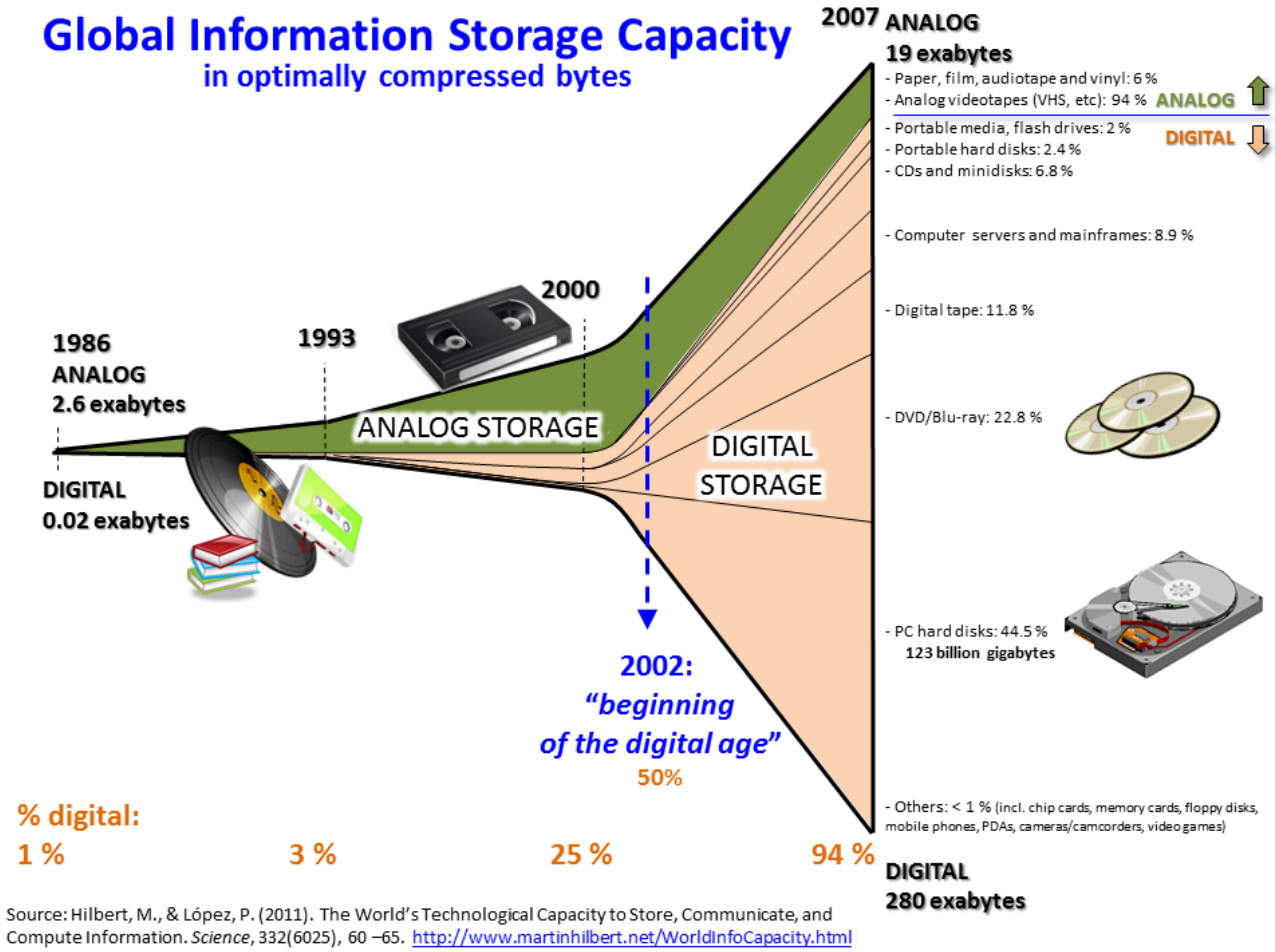
Why Big Data now?

Why did Big Data become hot now?

Big Data Challenge



2002: beginning of the Digital Age

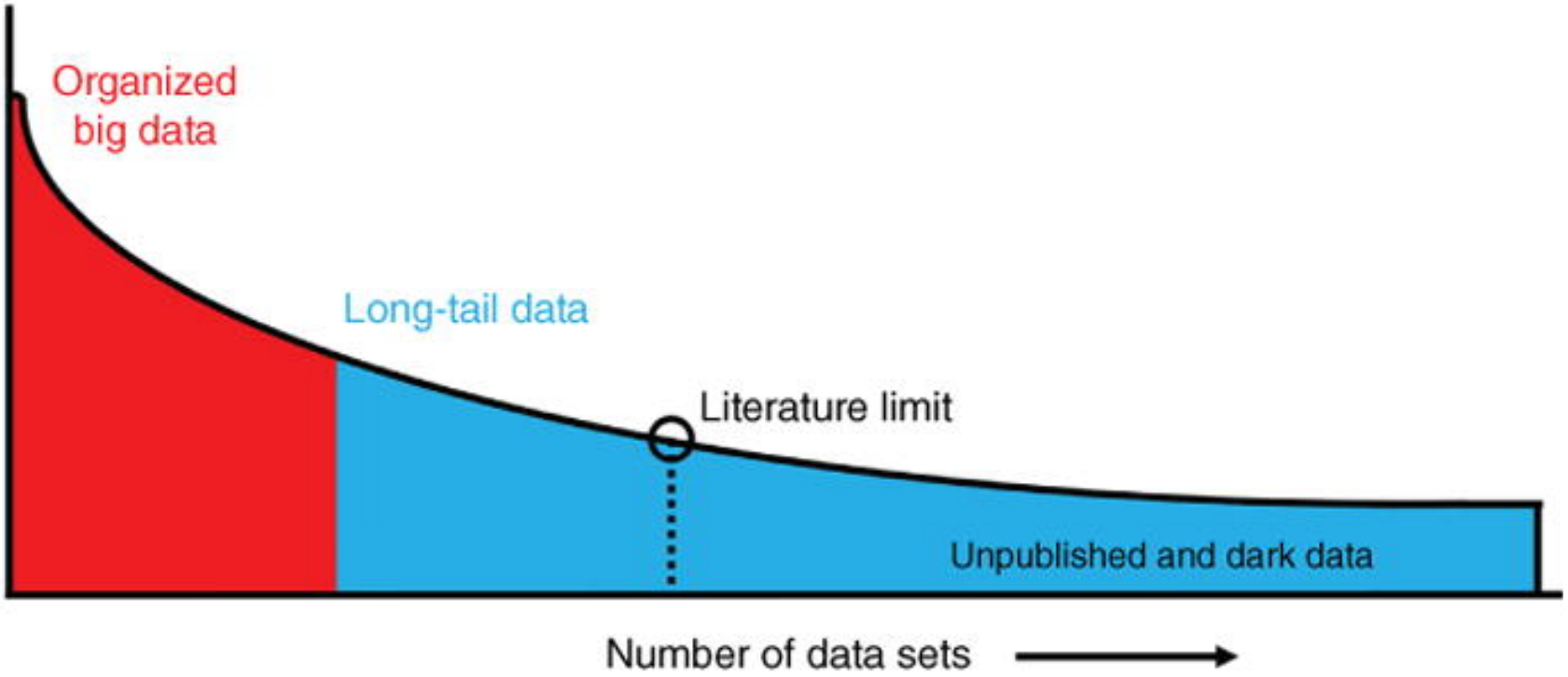


Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

- More data are being collected and stored;
- Open source code;
- Commodity hardware.

Science, 332(6050), 2011

Long-tail Data



- **Organized big data**. Well-organized big science efforts featuring homogenous, well-organized data represent *only a small proportion of the total data collected by scientists*.

- **Unpublished & dark data**. A very large proportion of scientific data falls in the long-tail of the distribution, with numerous small independent research efforts yielding a rich variety of specialty research data sets. The extreme right portion of the long tail includes data that are unpublished; such as siloed databases, null findings, laboratory notes, animal care records, etc. *These dark data hold a potential wealth of knowledge but are often inaccessible to the outside world*.

Nature Neurosciene, 17(11):1442, 2014

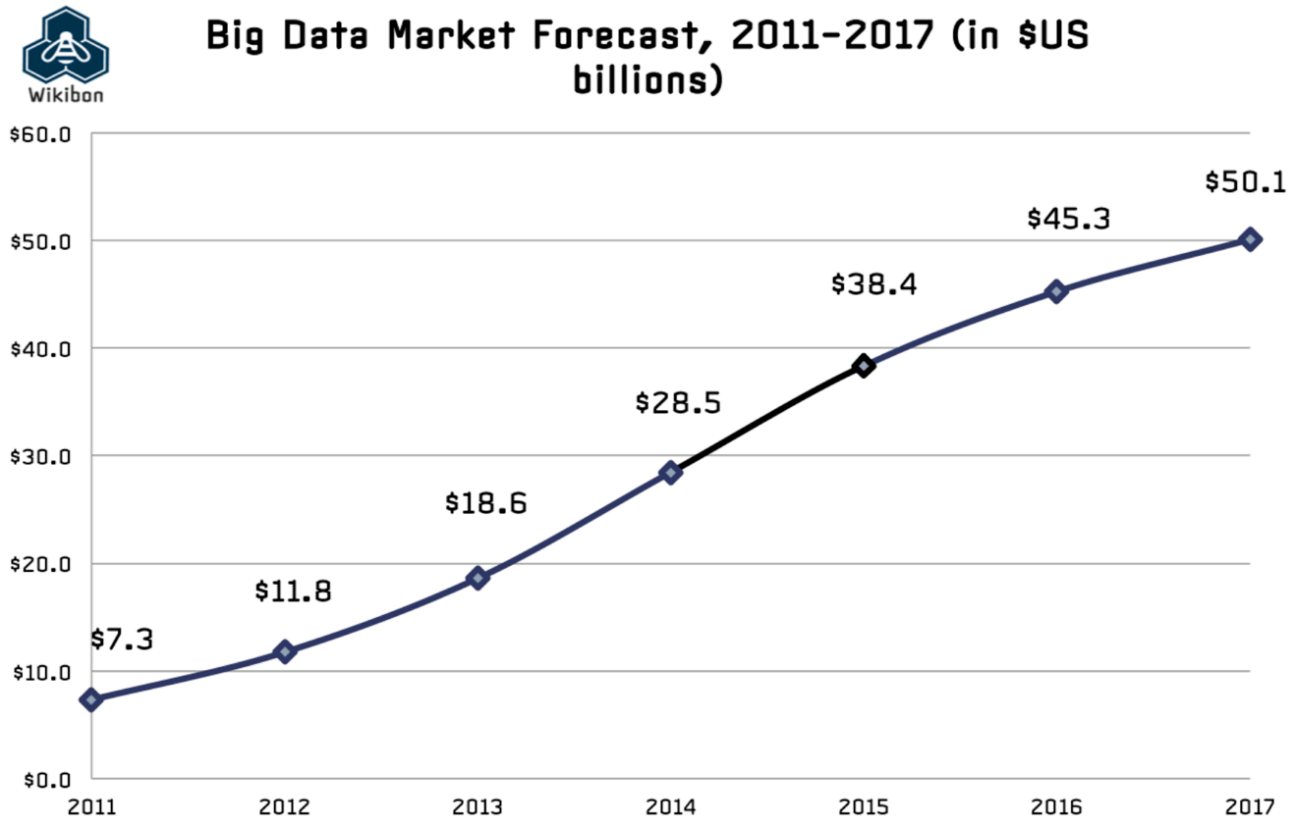
Techniques towards Big Data

- Massive Parallelism
- Huge Data Volumes Storage
- Data Distribution
- High-Speed Networks
- High-Performance Computing
- Task and Thread Management
- Data Mining and Analytics
- Data Retrieval
- Machine Learning
- Data Visualization

What made Big Data needed?

- Increased data volumes
- Increased computation need
- Increased Analytics need
- Cost Effective
- Innovative Techniques
- Lowered barrier to entry and success

"Big Data Analytics", David Loshin, 2013



http://wikibon.org/wiki/v/Big_Data_Vendor_Revenue_and_Market_Forecast_2013-2017

Big Data Market further breakdown

The "**Big in Big Data"** are:

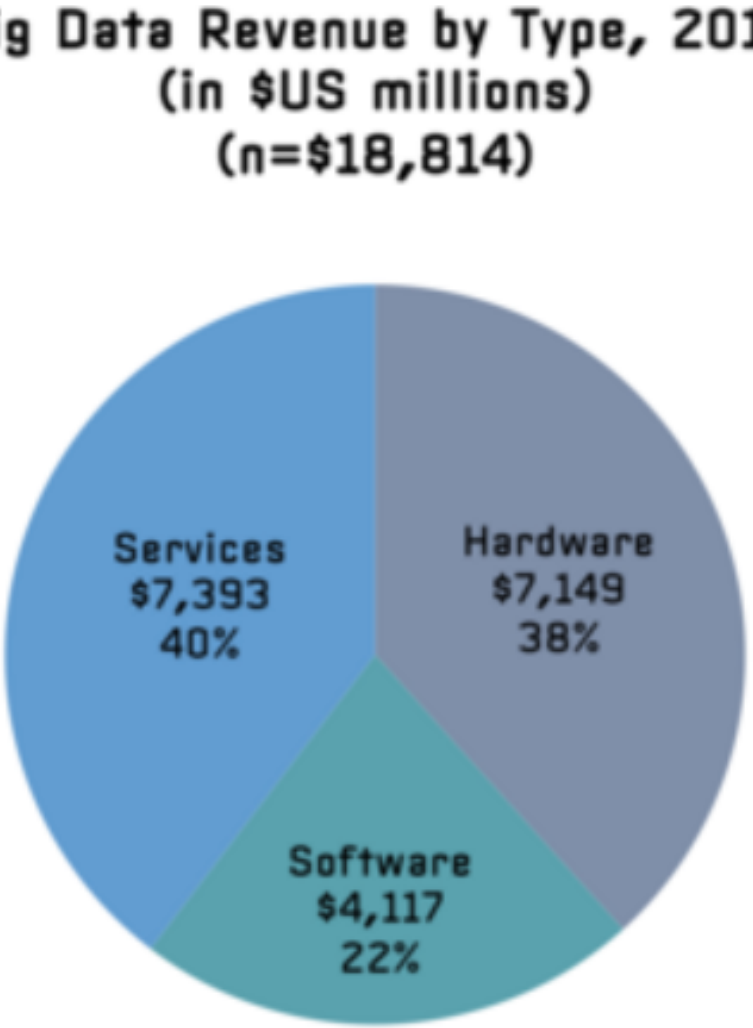
- Professional Services
- Application (Analytics)
- Storage
- Computation

	2014	2015	2016	2017
Big Data XaaS Revenue	$1.71	$2.43	$2.87	$3.19
Big Data Professional Services Revenue	$9.24	$12.31	$14.06	$15.30
Big Data Application (Analytic and Transactional) Revenue	$3.24	$4.94	$6.05	$6.89
Big Data NoSQL Database Revenue	$0.73	$1.14	$1.41	$1.62
Big Data SQL Database Revenue	$2.00	$2.48	$2.74	$2.91
Big Data Infrastructure Revenue	$0.67	$0.93	$1.08	$1.19
Big Data Networking Revenue	$0.67	$0.89	$1.02	$1.11
Big Data Storage Revenue	$4.39	$5.85	$6.68	$7.27
Big Data Compute Revenue	$5.23	$6.70	$7.50	$8.06
Total Big Data Revenue	$27.9	$37.7	$43.4	$47.5

USD: billions

http://wikibon.org/wiki/v/Big_Data_Vendor_Revenue_and_Market_Forecast_2013-2017

Big Data Market: Services, HW & SW



http://wikibon.org/wiki/v/Big_Data_Vendor_Revenue_and_Market_Forecast_2013-2017

Big Data Market

2013 Worldwide Big Data Revenue by Vendor ($US millions)

Vendor	Big Data Revenue	Total Revenue	Big Data Revenue as % of Total Revenue	% Big Data Hardware Revenue	% Big Data Software Revenue	% Big Data Services Revenue
IBM	$1,368	$99,751	1%	31%	27%	42%
HP	$869	$114,100	1%	42%	14%	44%
Dell	$652	$54,550	1%	85%	0%	15%
SAP	$545	$22,900	2%	0%	76%	24%
Teradata	$518	$2,665	19%	36%	30%	34%
Oracle	$491	$37,552	1%	28%	37%	36%
SAS Institute	$480	$3,020	16%	0%	68%	32%
Palantir	$418	$418	100%	0%	50%	50%
Accenture	$415	$30,606	1%	0%	0%	100%
PWC	$312	$32,580	1%	0%	0%	100%
Deloitte	$305	$33,050	1%	0%	0%	100%
Pivotal	$300	$300	100%	15%	50%	35%
Cisco Systems	$295	$50,200	1%	72%	12%	16%

http://wikibon.org/wiki/v/Big_Data_Vendor_Revenue_and_Market_Forecast_2013-2017

Big Data: Astronomical or Genomical?

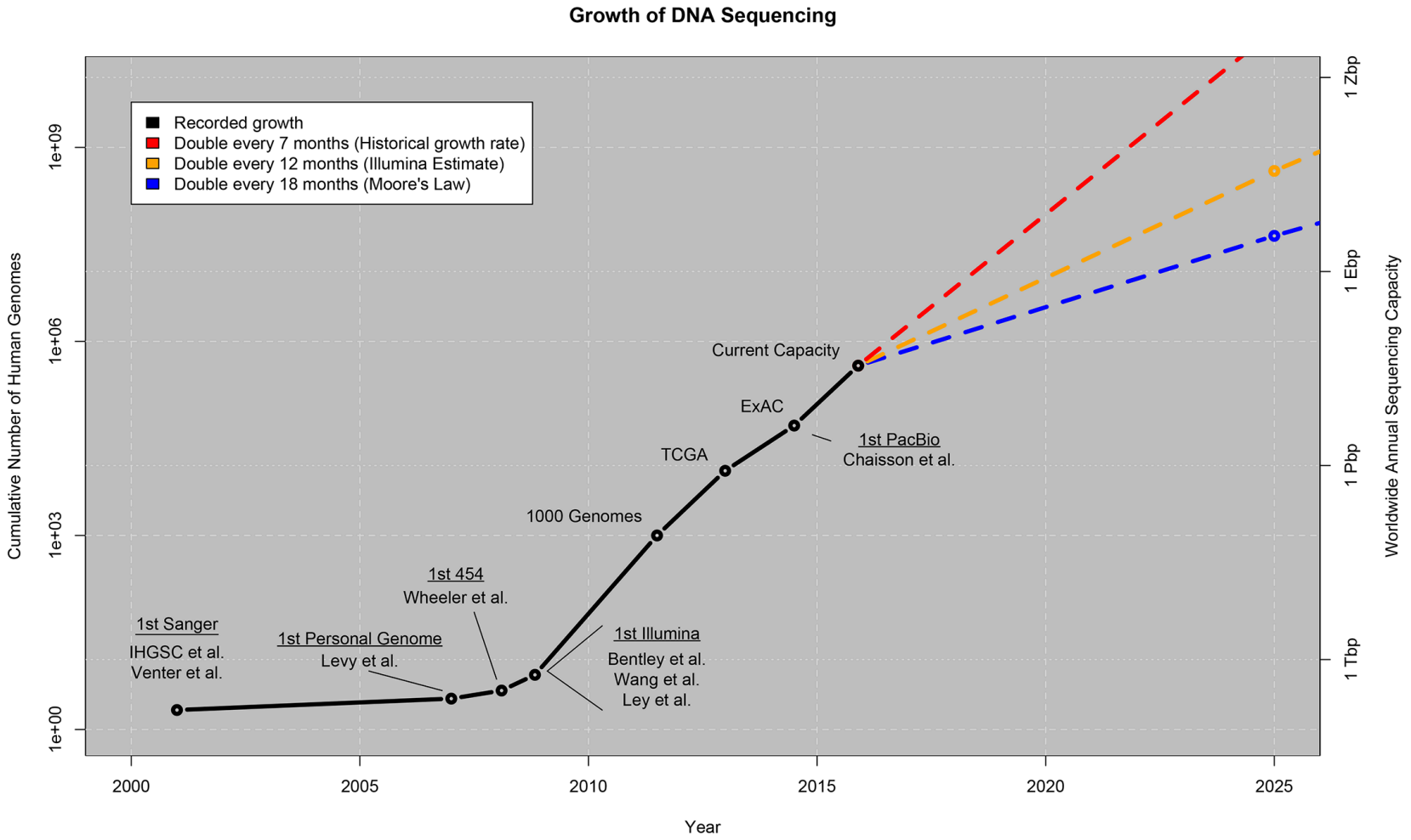
Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

doi:10.1371/journal.pbio.1002195.t001

Four domains of Big Data in 2025. In each of the four domains, the projected annual storage and computing needs are presented across the data lifecycle (Exa=2^{60}, Zetta = 2^{70}).

PLOS Biology, 13(7), 2015

Big Data: Astronomical or Genomical? Growth of DNA sequencing



Total number of human genomes sequenced (left axis) as well as the worldwide annual sequencing capacity (right axis: Tera-basepairs (Tbp), Peta-basepairs (Pbp), Exa-basepairs (Ebp), Zetta-basepairs (Zbps)).

PLOS Biology, 13(7), 2015

Analytics – Unlocking the Potential of Big Data

Descriptive Analytics – "*What is happening?*"

Querying, Reporting, Data Capturing, Filtering & Analysis

Predictive Analytics – "*What will likely happen?*"

Statistical Methods (Regression), Forecasting & Data Mining

Prescriptive Analytics – "*What should we do?*"

Optimization, Simulation & Quantitative Models

Big Data Analytics Example Use Cases

1. Social Network Analysis
2. Recommendation
3. Commerce
4. Financial Analysis
5. Social Media Monitoring
6. Customer Analysis
7. IBM Watson
8. Data Exploration and Visualization (Challenges by Convolutional NN)
9. Personalized Search
10. Anomaly Detection (Espionage, Sabotage, etc.)
11. Fraud Detection
12. Cybersecurity
13. Sensor Monitoring
14. Cellular Network Monitoring
15. Cloud Monitoring
16. Code Life Cycle Management
17. Traffic Navigation
18. Image and Video Semantic Understanding (e.g., **scene understanding**)
19. Genomic Medicine (e.g., **Genome-Write Project**)
20. Brain Network Analysis
21. Data Curation
22. Near Earth Object Analysis

Challenges & Open Problems

76425 species

Tree of Life by Dr. Yifan Hu

14.8 million tweets

The information diffusion graph of the death of Osama bin Laden by Gilad Lotan

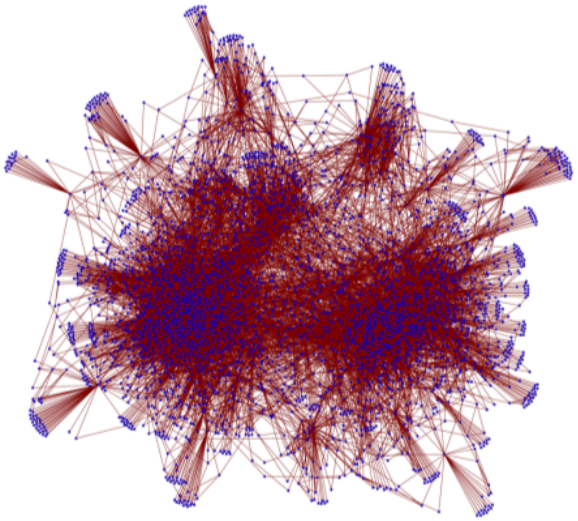
500 million users

Facebook friendship graph by Paul Butler

- **How to Visualize Huge Static Graph**:

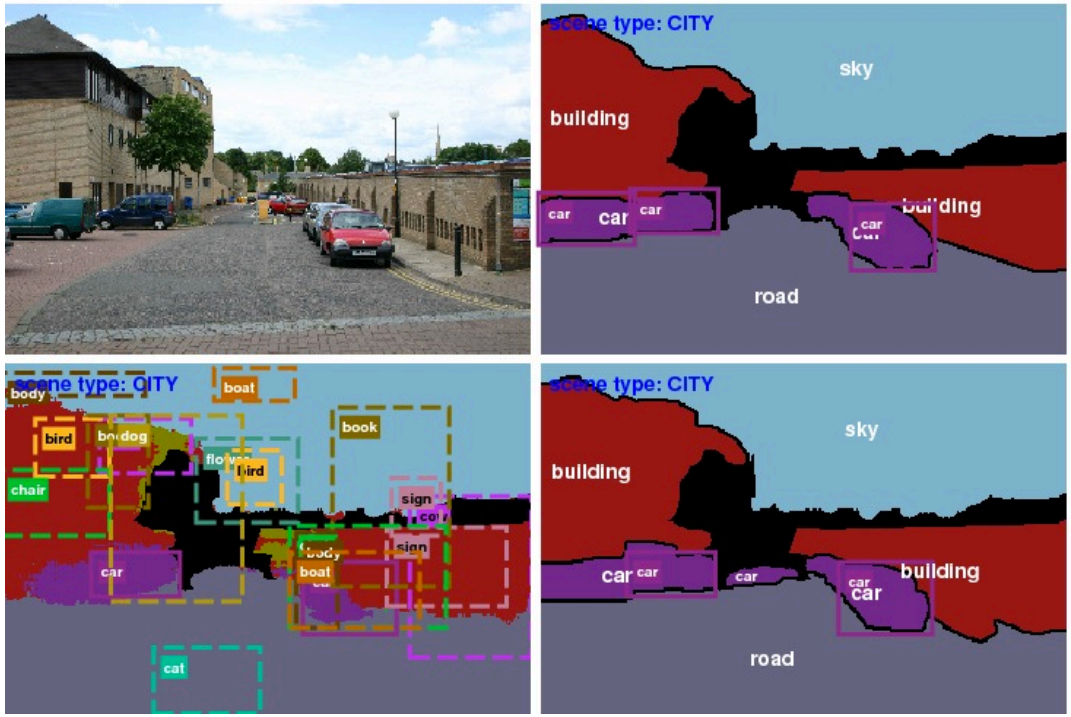
  Squeezing millions and even billions of records into million pixels (1600 X 1200 ≈ 2 million pixels)

Visualization Key Challenges



- **Visual clutter**
  - How can we *encode the information intuitively*?
- **Performance issues**
  - How can we *render the huge datasets in real time with rich interactions*?
- **Cognition**
  - *How can users understand the visual representation* when the information is overwhelming?

Towards Scene Understanding: Object Detection and Contextual Reasoning



Scene understanding is one of the holy grails of computer vision. Despite decades of research on scene understanding, it is still considered an unsolved problem. The difficulty arises mainly because of the huge space of possible images. We require models to capture this variability of scenes and their constituents (e.g., objects) given the limited memory resources. Additionally, we require efficient learning and inference techniques for our models to find the optimal solution in the enormous space of possible solutions.

Challenges in big data

- **Challenges in big data applications**
  - The problem of dimensionality
  - Storage cost
  - Query speed
- **Why Learning from Big Data is Challenging?**
  - High per-iteration cost
  - High memory cost
  - High communication cost (e.g., cloud - mobile)
  - Large iteration

Seminal References

- Nature, special issue on Big Data, 2008
- Science, special issue on Data, 2011
- "*The Fourth Paradigm - Data-Intensive Scientific Discovery*", edited by Tony Hey, Stewart Tansley & Krist in Tolle, 2009
- "*Big Data Analytics*", David Loshin, Morgan Kaufmann, 2013

Nature, 3 Set 2008

Science, 11 Feb 2011

Journals on Big Data

- Big Data Analytics – BioMed Central (2016-)
- IEEE Transactions on Big Data (2015-)
- Journal of Big Data – Springer (2014-)
- Big Data Research – Elsevier (2013-)
- Big Data - Mary Ann Liebert, Inc. (2013-)
- …

Analytics – Unlocking the Potential of Big Data

Descriptive Analytics – "*What is happening?*"

Querying, Reporting, Data Capturing, Filtering & Analysis

Predictive Analytics – "*What will likely happen?*"

Statistical Methods (Regression), Forecasting & Data Mining

Prescriptive Analytics – "*What should we do?*"

Optimization, Simulation & Quantitative Models

Prescriptive Big Data Analytics by Pareto Optimality

"*What should we do?*"

Solar Cells, Airline Transportation Networks, Electrics Networks, Transistors Networks, Amino Acids Networks and Metabolic Networks

Giuseppe Nicosia

Dept of Mathematics & Computer Science, University of Catania, Italy

www.dmi.unict.it./nicosia/

Agenda

- **MOO**
- **Algorithms and Methods for designing high performance** systems of systems
- **Solar Cells**
- **Airline Transportation Systems as Multiplex Networks**
- **Semiconductor Design**
  - *MESFETs, MOSFETs , Double Gate MOSFETs*
- **Circuit Design**
  - RF Low Noise Amplifier 3-5GHz
  - Leapfrog Filter Ultra Wideband
  - Fully Differential Folded Cascode Operational Amplifiers @ 3 temperatures
- **BioPlastic production using Yeast**
- **Protein Structure Prediction**
- **Conclusions**

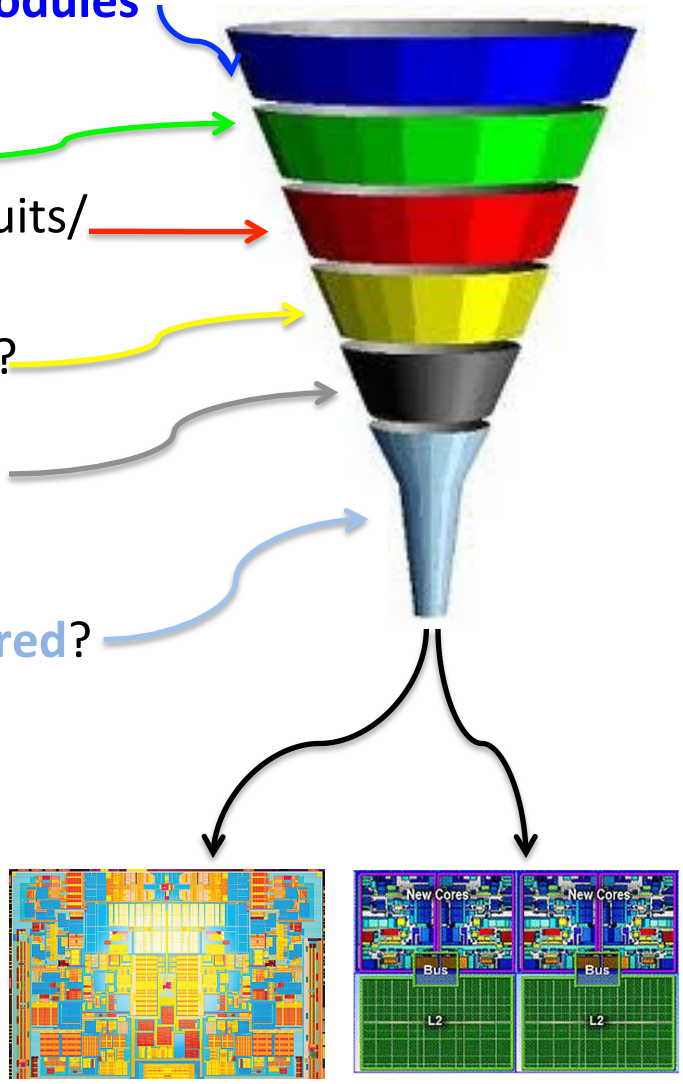
Pareto Fronts as Power Laws

A wide variety of physical, chemical, biological, and man-made phenomena/systems approximately show characteristic Pareto fronts.

- **local (bond atoms) vs. non-local (non-bond atoms)** in the proteins (*networks of amino-acids*),
- **CO_2 Uptake vs. Nitrogen consumption** in the photosynthesis of chloroplasts (*networks of enzymes*),
- **Succinate/Acetate vs. biomass production** in E. coli, Ethanol and biomass production in E. coli (*networks of metabolites*)
- **Lactate vs. biomass production** in S. cerevisiae (*networks of gene sets, fluxes and metabolites*)
- **current consumption vs. noise** in analog circuits (*network of transistors*),
- **quantum efficiency vs. thickness** in solar cells (*networks of Si atoms*),
- **power consumption vs. miss in cache** in CPUs (*networks of circuits and systems*).

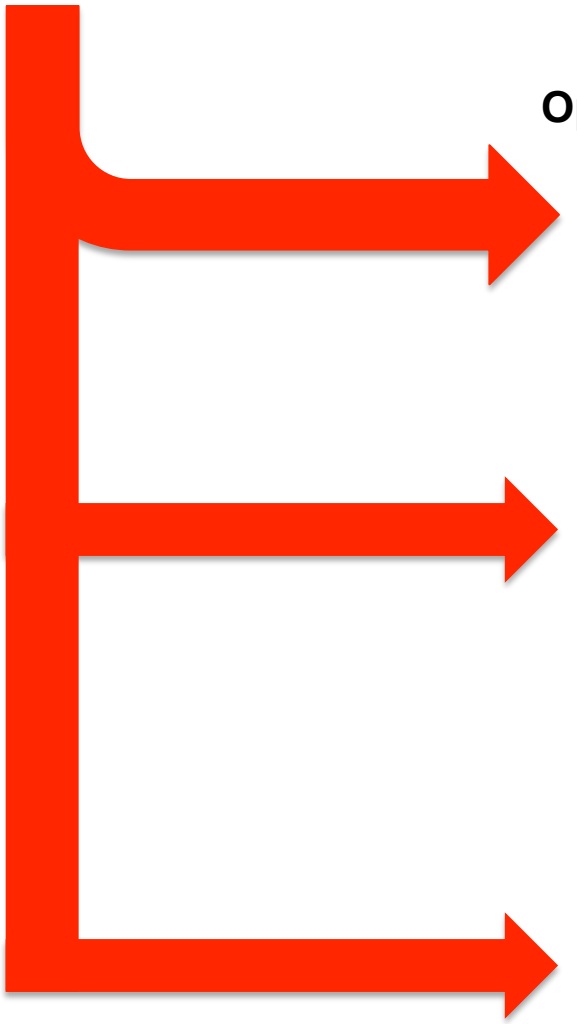
Multi-objective Multi-criteria Optimization Methodology

- Which are **the most important parameters/parts/modules** of the given Device/Circuit/System?
- How many **Feasible** Devices/Circuits/Systems?
- How many **Optimal and/or Suboptimal** Devices/Circuits/Systems?
- How many **Pareto Optimal** Devices/Circuits/Systems?
- How many **Robust Pareto Optimal** Devices/Circuits/Systems?
- Which is the set of **Robust Pareto Optimal** Devices/Circuits/Systems that can be **successfully manufactured**?



Optimization in one Slide

- **Optimization problems**
  - Uni- or Multi- modal
  - Local or Global
  - Multi-objective
  - Combinatorial
  - Nonlinear
  - Convex / NoConvex
  - Constrained
  - Mixed
  - …
- **Parameter types**
  - Continuous
  - Binary/Discrete
  - Integer
  - Num. Par. (10-10^6)
  - Num. Objs/Con. (1-10^2)
  - Mixed
  - …
- **Search spaces**
  - Rugged
  - Smooth
  - Funnel
  - Huge/Massive
  - …



Optimization Algorithms ?

- Deterministic
- Randomized
- Derivative-Free
- Black-Box
- Nature-Inspired
- Integer Programming
- Mathematical Programming
- Hybrid approach
- Geometric Programming
- Commercial tools
- To develop new algorithms
- Examples: DIRECT, CMA-ES, GA, DE, ES, GP, CRS, SA, IA, GPS, MADS, PAES, Nelder-Mead Simplex, SPEA, NSGA-II, Trust-Region, PSO, etc. etc.
- …

Constrained Multi-Objective Optimization

- **MOO**: the approach is motivated by the observation that most EDA problems involve
  - ***multiple***,
  - ***conflicting***, and
  - ***non-commensurate objectives (goals)***.

    J. ACM, 38(4), 1991, 775-814.
  - ***And (strong) nonlinear interdepence.***

- **Constraints**: optimizing **multiple conflicting objectives** while satisfying several **constraints.**

- **Problem domain**: Variables are commonly defined in ***a mixed integer/ discrete/continuous domain. In particular,***

[Li, Ui, Si]

Device/Circuit/System Design Problems can be defined as a ***constrained multi-objective optimization problem*** defined in ***a mixed integer/discrete/ continuous domain***.

Objectives, Constraints & Variables

- Typical ***Objectives and Constraints targets*** used in EDA design are
  - Silicon area,
  - noise,
  - power consumption,
  - current gain,
  - θ deviation from the reference doping profile,
  - energy, etc.
- As function of ***Design Variables*** such as
  - Transistor sizes (width and length),
  - Resistor/capacitor values,
  - doping concentration c_0 in the n^+ regions,
  - doping concentration c_1 in the n region,
  - L1 Instruction Cache size, etc

Mixed Integer-Discrete-Continuous Constrained MOOP

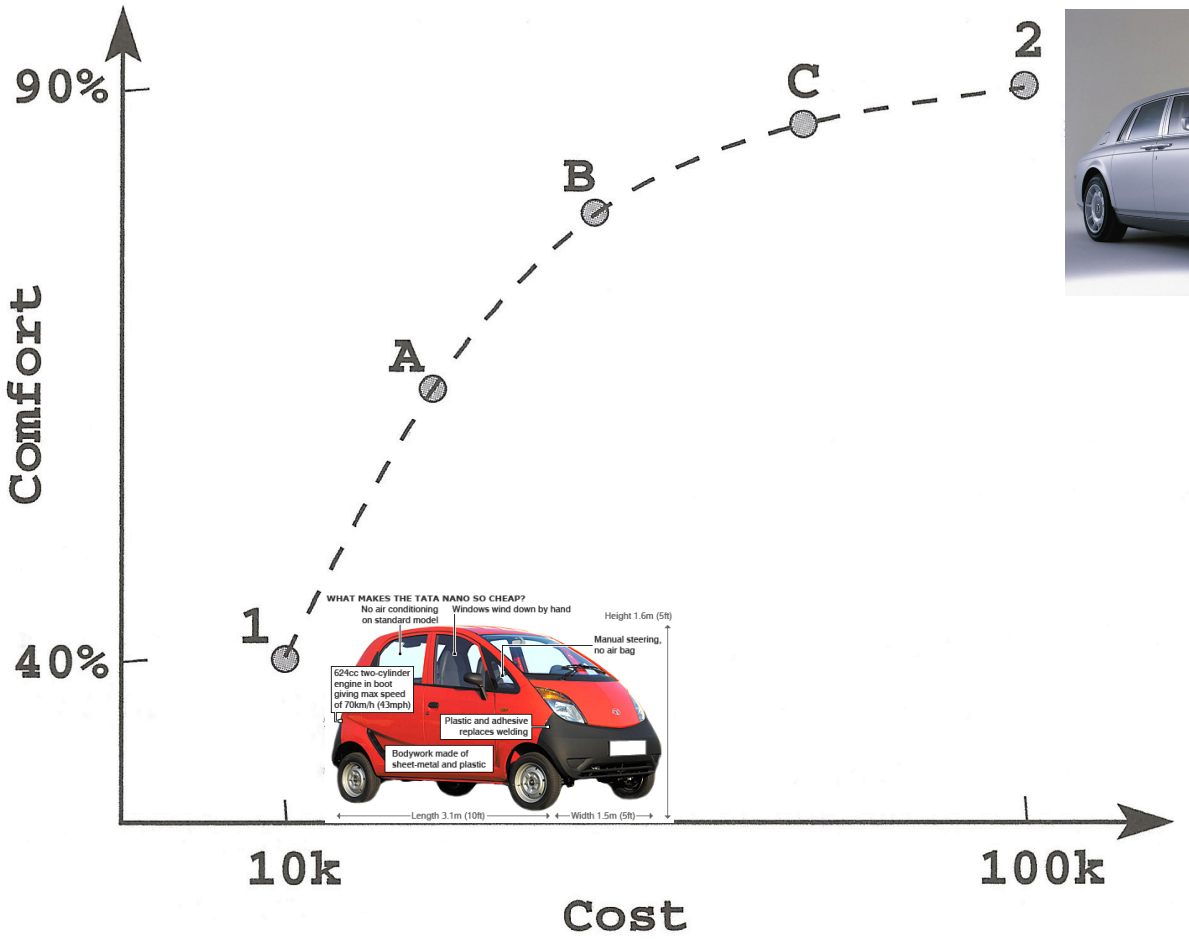
Find	$X=\{x_1, x_2, \ldots, x_n\} = [X^{(i)}, X^{(d)}, X^{(c)}]^T$	
To Minimize/maximize	$f_m(x)$,	$m = 1,2,\ldots,\mathbf{M}$;
Subject to	$g_j(x)\geq 0$,	$j = 1,2,\ldots,J$;
	$h_k(x)=0$,	$k = 1,2,\ldots,K$;
	$x_i^{(L)} \leq x_i \leq x_i^{(U)}$	$i = 1,\ldots,N$.

Where $X^{(i)}$, $X^{(d)}$, $X^{(c)}$ denotes **feasible subsets of integer, discrete and continuous variables respectively**. While both integer and discrete variables have a discrete nature, only discrete variables can assume floating point values (*they are often unevenly spaced*): **[Li, Ui, Si]**

Integer and discrete variables required different handling.

If a solution x satisfies all of the **(J+K) constraints** and all of the **2N variable bounds**, it is known as a **feasible solution**.

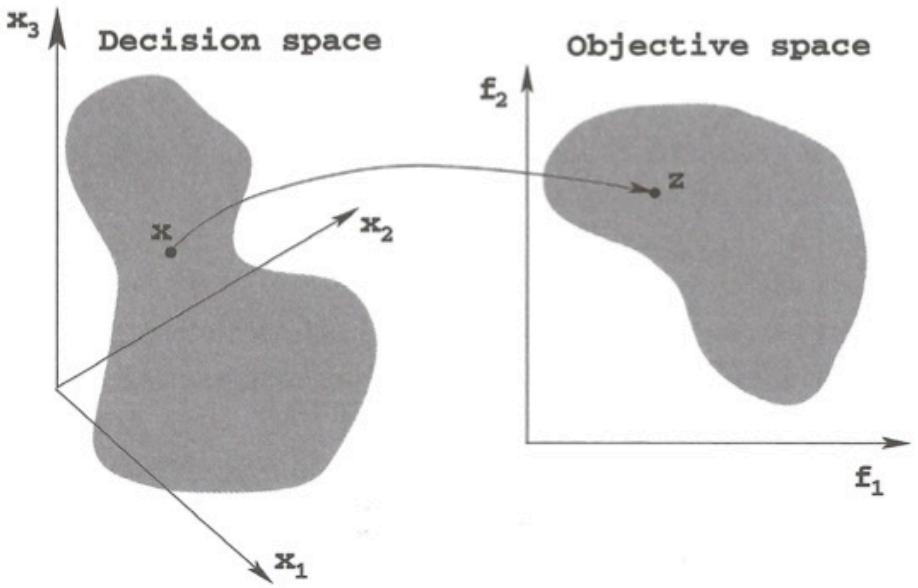
Why MOO?

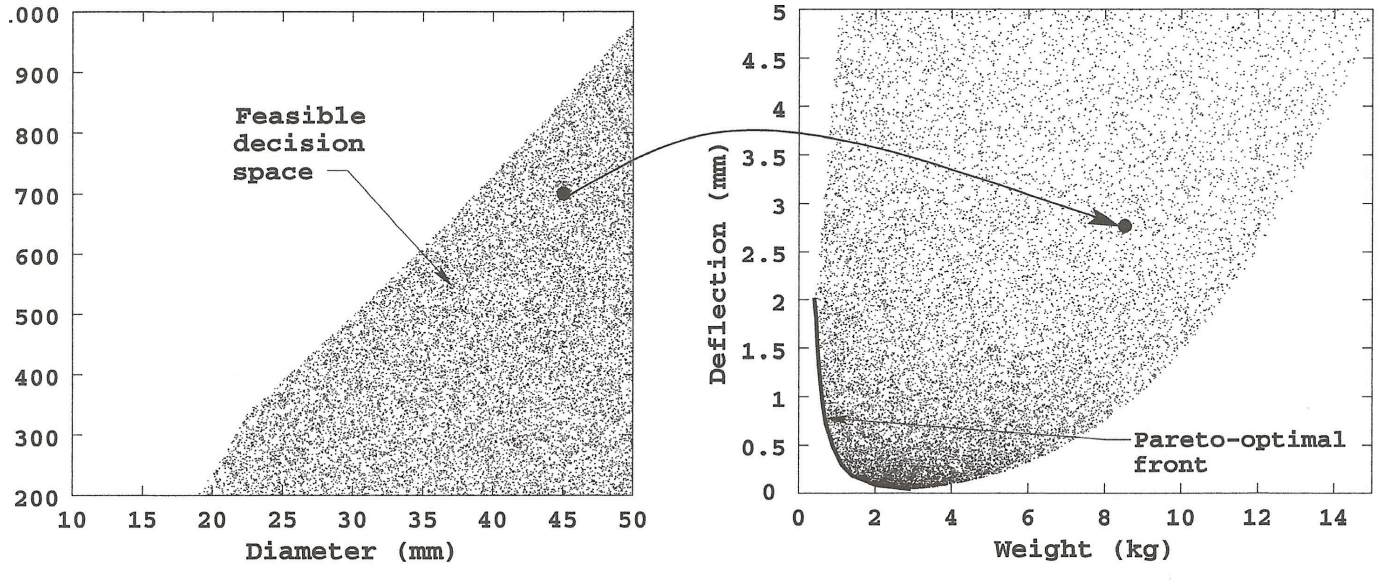


Conflicting objectives: The Vilfredo Pareto's compromises.

Credits: K. Deb.

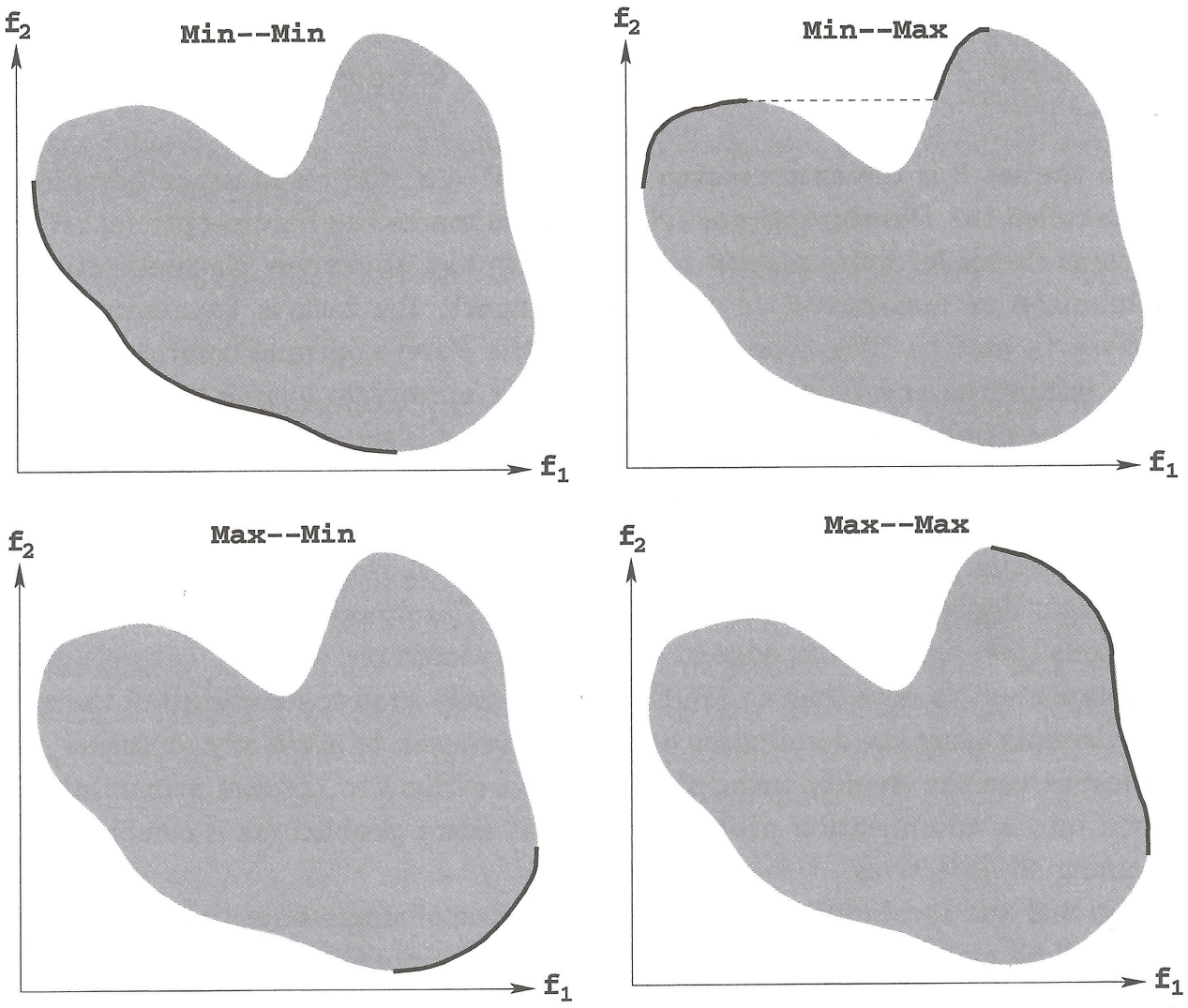
Decision Space vs. Objective Space





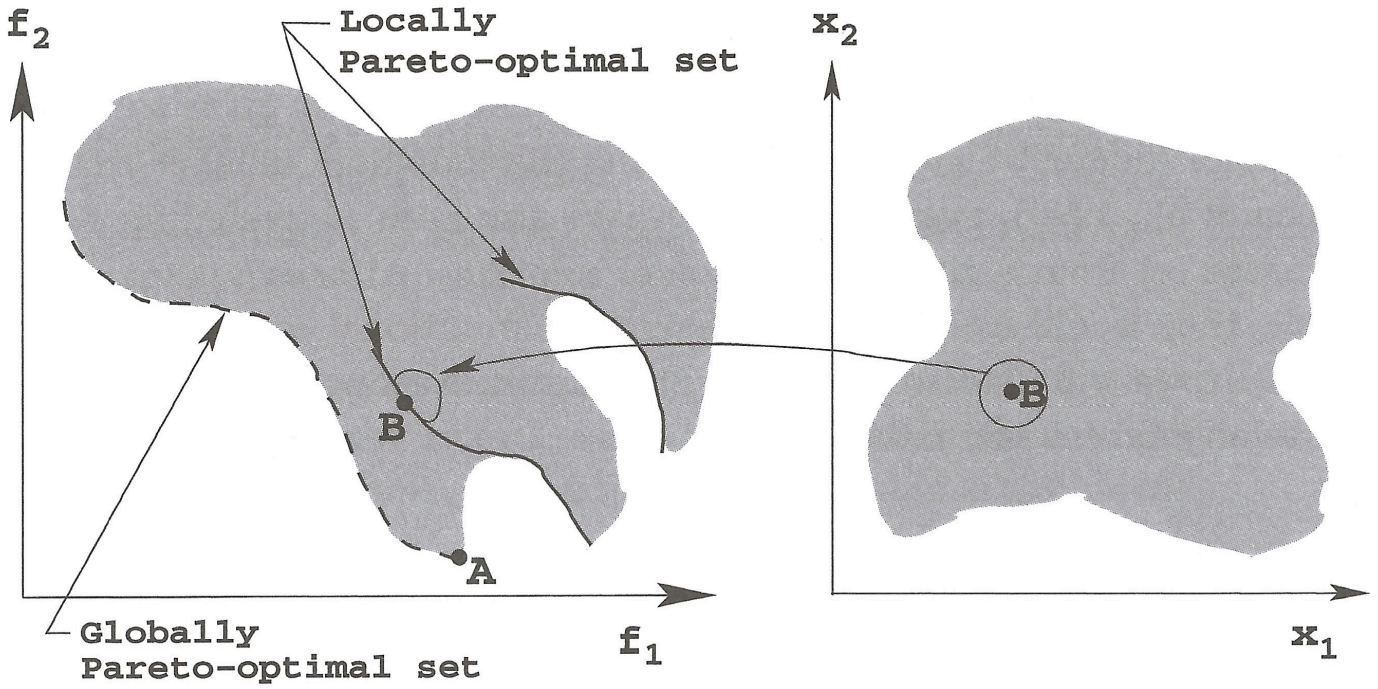
Credits: K. Deb.

4 Objective Spaces for 2 Objective Optimization



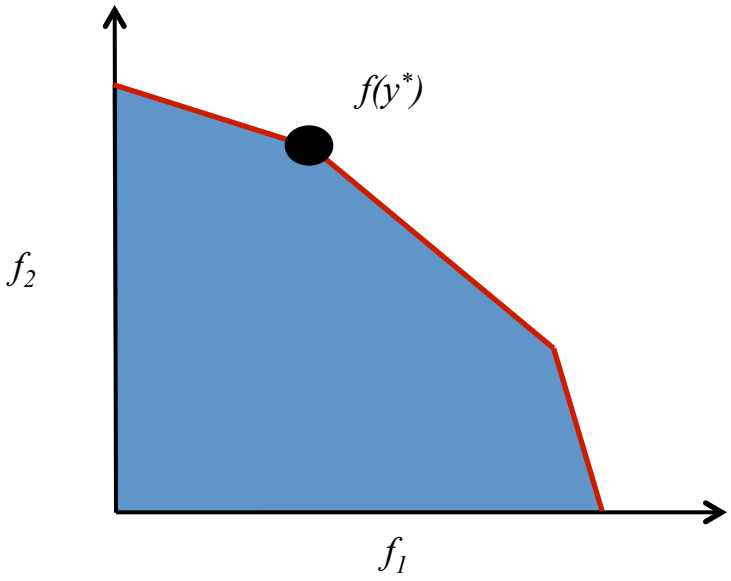
Credits: K. Deb.

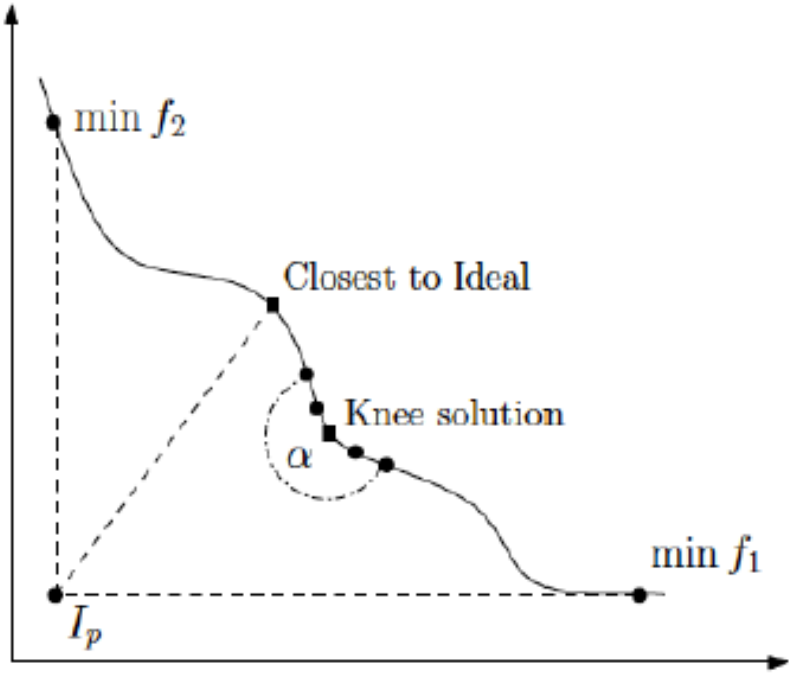
Locally vs. Globally Pareto Optimal Set



Credits: K. Deb.

Pareto Fronts (PF) & Decision Making

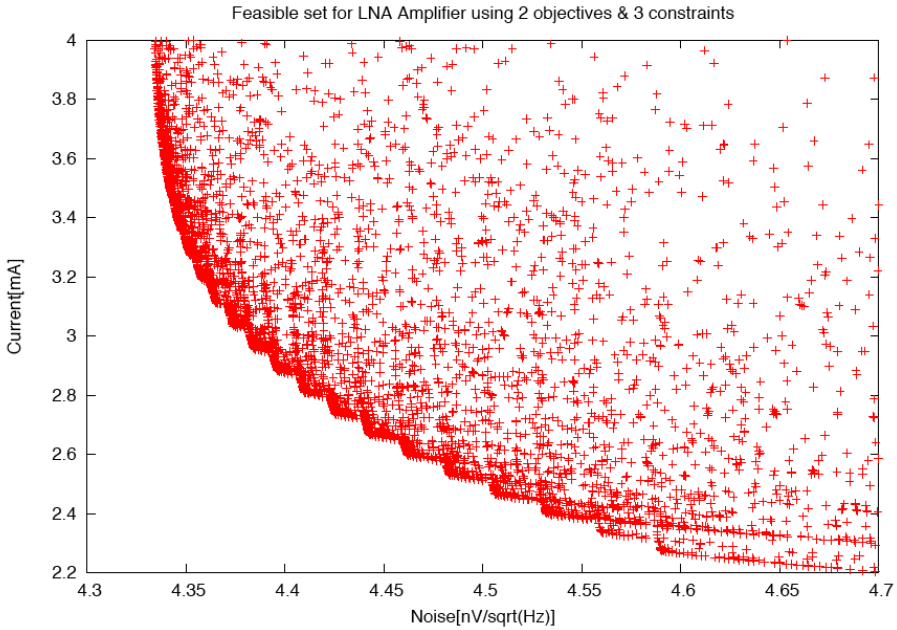




Decision making strategies. Geometrical representation of the various strategies on a bi-objective Pareto front.

Pareto Optimum (Vilfredo Pareto 1896)

x* is Pareto optimal if there exists no feasible point x which would ***decrease some criterion without causing a simultaneous increase in at least one other criterion***.



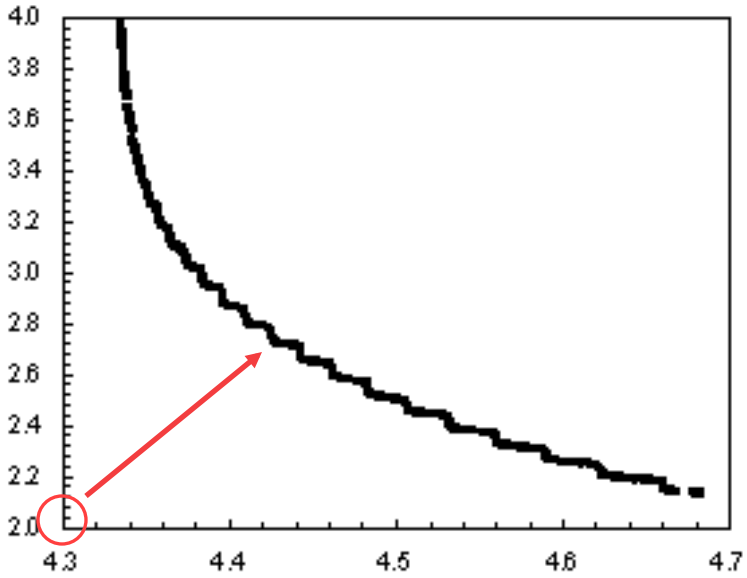
The notion of "optimum" changes: The aim is to find good compromises (or trade-offs) *rather than a single solution as in global optimization*.

1. To find a $PF_{observed}$ ***as close as possible to PF****
2. To find a $PF_{observed}$ ***as diverse as possible***

Decision Making Phase

- For a MOOP we can define the following procedure:
  1. Find the optimal (or the observed) Pareto front; and
  2. Choose one of the candidate solutions in the Pareto front, using some higher-level information.

- ***The selection criterion***: to select the non-dominated solution closest to the ***ideal point*** (for each k objectives there exists one different optimal solution; an objective vector constructed with these individual optimal objective values constitutes the ***ideal objective vector***).



Higher-level information for CDP : **yield, robustness, costs, expected Time-to-market**, etc.

The optimization Algorithm

A Population-based Stochastic Generate-and-Test Algorithm

The optimization Algorithm

Yet Another Population-based Stochastic Generate-and-Test Algorithm

opt-IA $(d, dup, \tau_B, \rho, \beta)$
1. $t := 0$
2. $P^{(t)} :=$ Initial_Pop()
3. Evaluate$(P^{(0)})$
4. **while** ($\neg$ Termination_Condition())*do*
5. $\quad P^{(clo)} :=$ Cloning $(P^{(t)}, dup)$
6. $\quad P^{(hyp)} :=$ Hypermutation$(P^{(clo)}, \rho)$
7. $\quad P^{(macro)} :=$ Macromutation$(P^{(hyp)}, \beta)$
8. $\quad$ Evaluate$(P^{(macro)})$
9. $\quad$ Aging$(P^{(t)}, P^{(macro)}, \tau_B)$
10. $\quad P^{(t+1)} := (\mu + \lambda)$-Selection$(P^{(t)}, P^{(macro)})$
11. $\quad t := t + 1$
12. **end_while**

G. Nicosia et al, J. Global Optimization, 53:4, 2012.

G. Nicosia et al., IEEE Trans Biomedical Circuits and Systems, 2016.

Hypermutations & Constraints

- The **hyper-mutation operator** mutates a randomly chosen variable x_i of a given candidate solution **x** using a *self-adaptive Gaussian mutation* computed as

$$x_i^{new} = x_i + \sigma N(0,1)$$

- the **hyper-macromutation** applies a **convex perturbation** to a given solution by setting

$$x_i^{new} = (1-\gamma)x_i + \gamma x_k$$

- These mutation operators are controlled by **specific mutation rates**; *for the hyper-mutation,* we define

$$\alpha = e^{-\rho f}$$

- instead *for the hyper-macromutation* we adopted

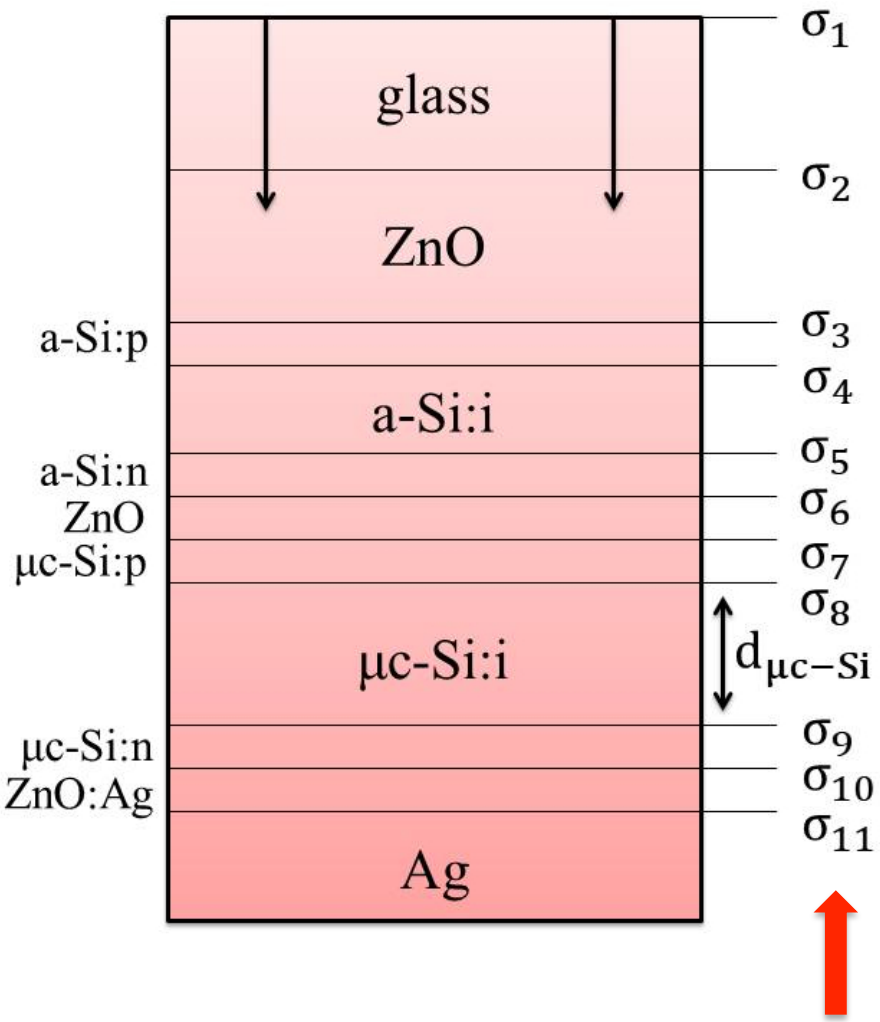
$$\alpha = \frac{1}{\beta} e^{-f}$$

- The algorithm considers the constraint values during the selection procedure. Given two individuals p_1,p_2,
  - **If** p_1,p_2 both are feasible **then** the one with the lowest objective function value is picked;
  - **If** p_1 is feasible and the p_2 is unfeasible **then** p_1 is chosen, otherwise
  - **If** p_1 and p_2 are unfeasible **then** the one with the lowest constraints violation is selected.

Solar Cells

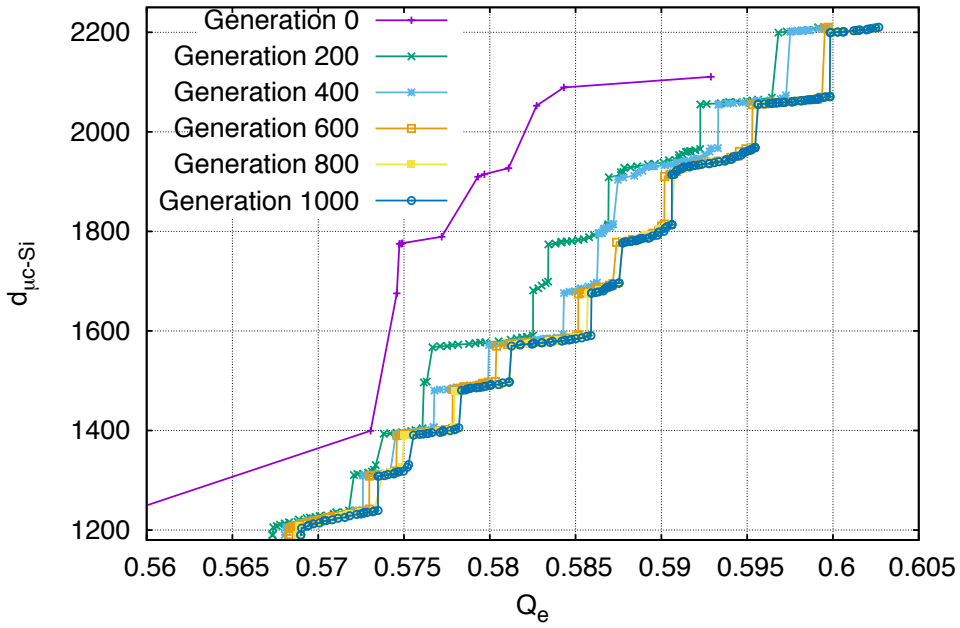
2-D Tandem Thin-Film Silicon Solar Cell Structure

- ZnO (**Zinc Oxide**) as TCO (**Transparent Conductive Oxide**)
- Ag (**Silver**) as BR (**Back Reflector**).

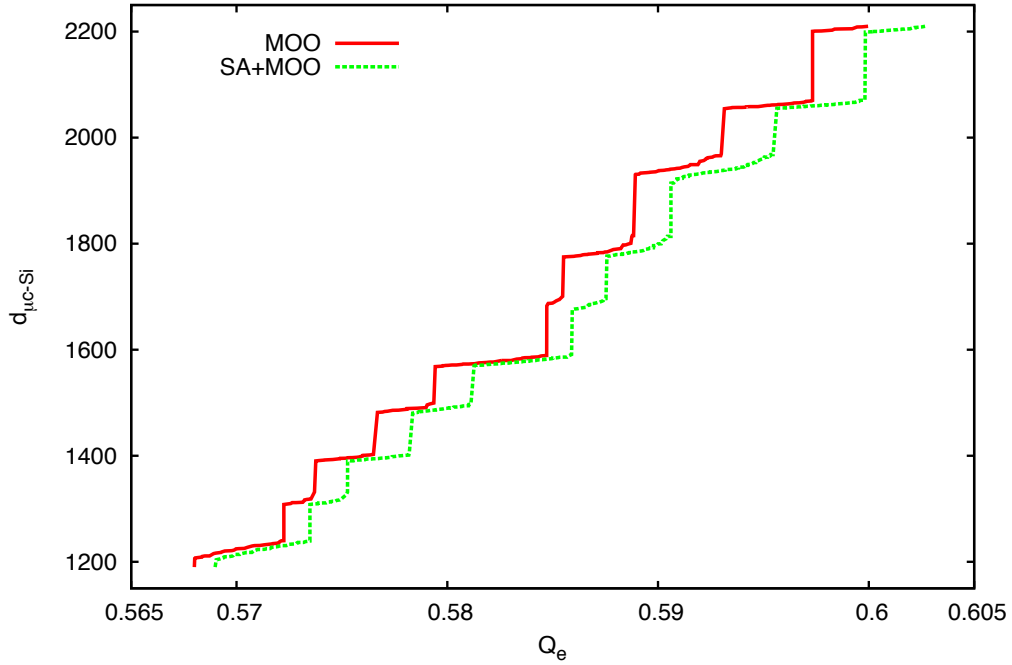


the parameters selected for the optimization process.

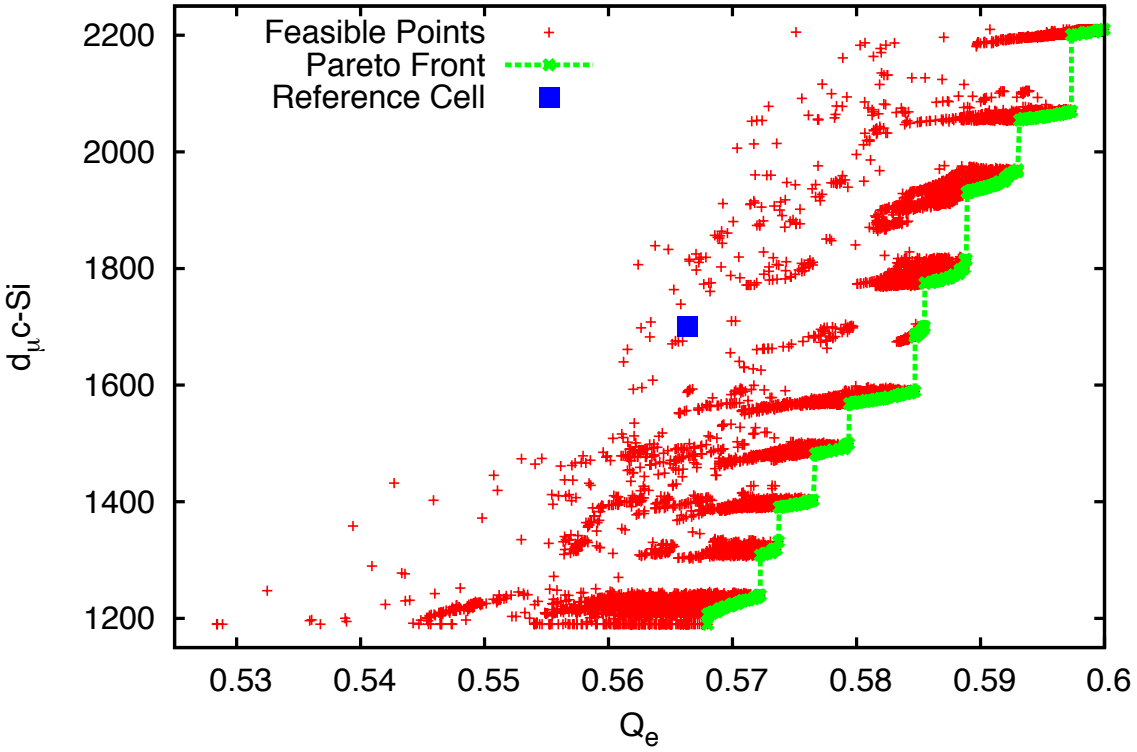
Multi-Objective Convergence Process

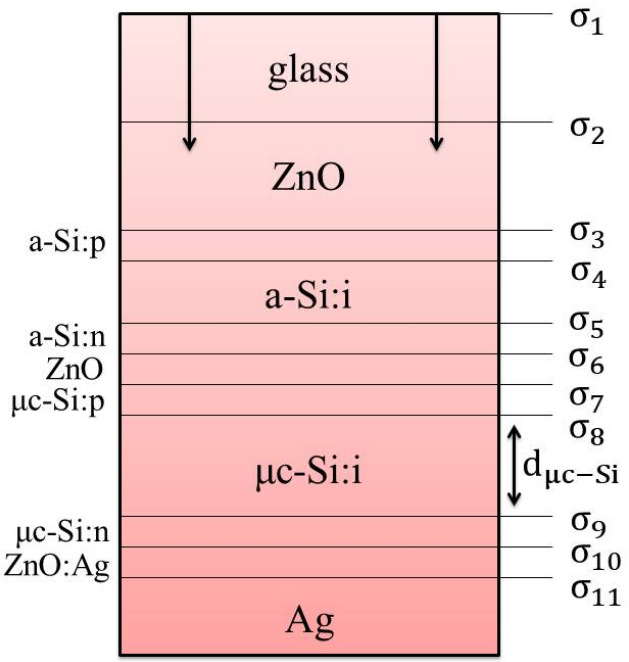


Speed-up Convergence Process

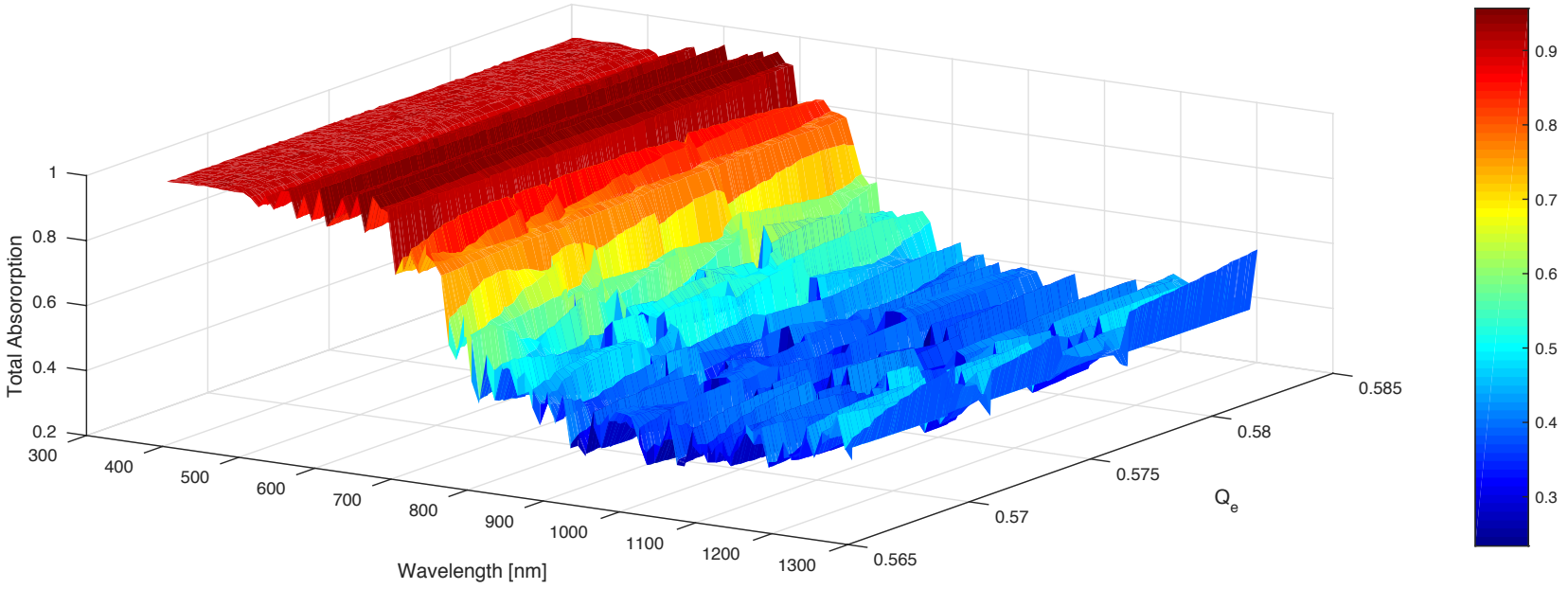


The Pareto Optimal solar cells



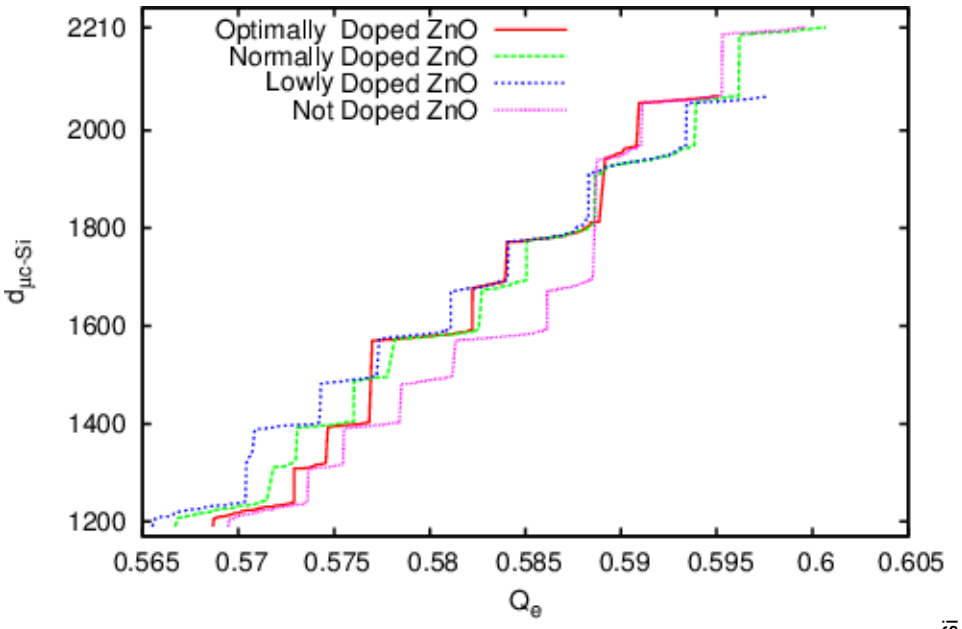


Pareto Optimal absorption Profiles

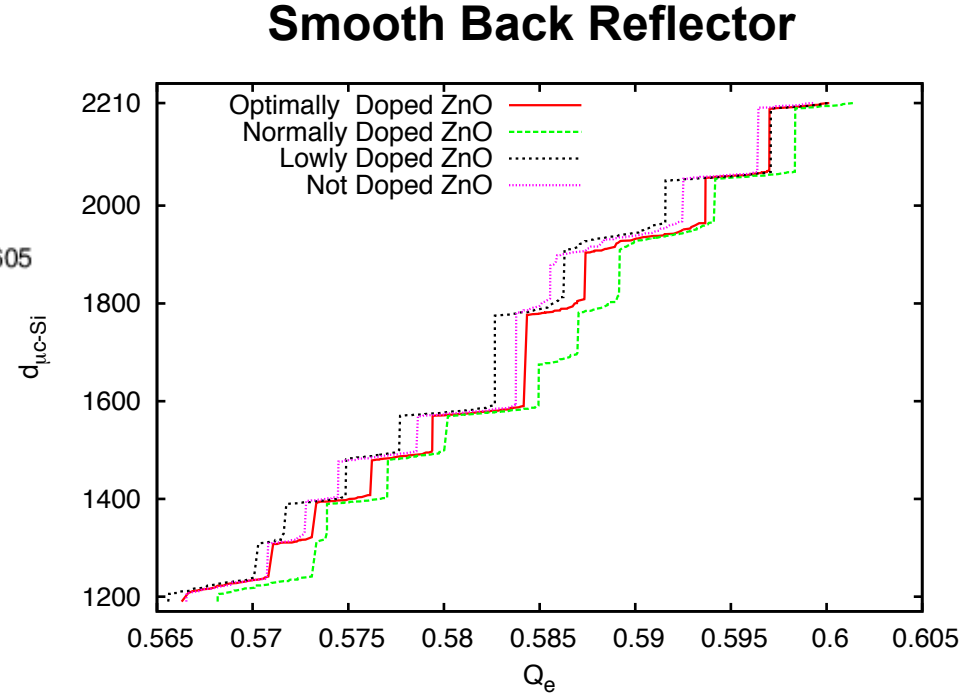


Studying Materials with Pareto Optimality
Different Doping Dosage and Roughness

Rough Back Reflector



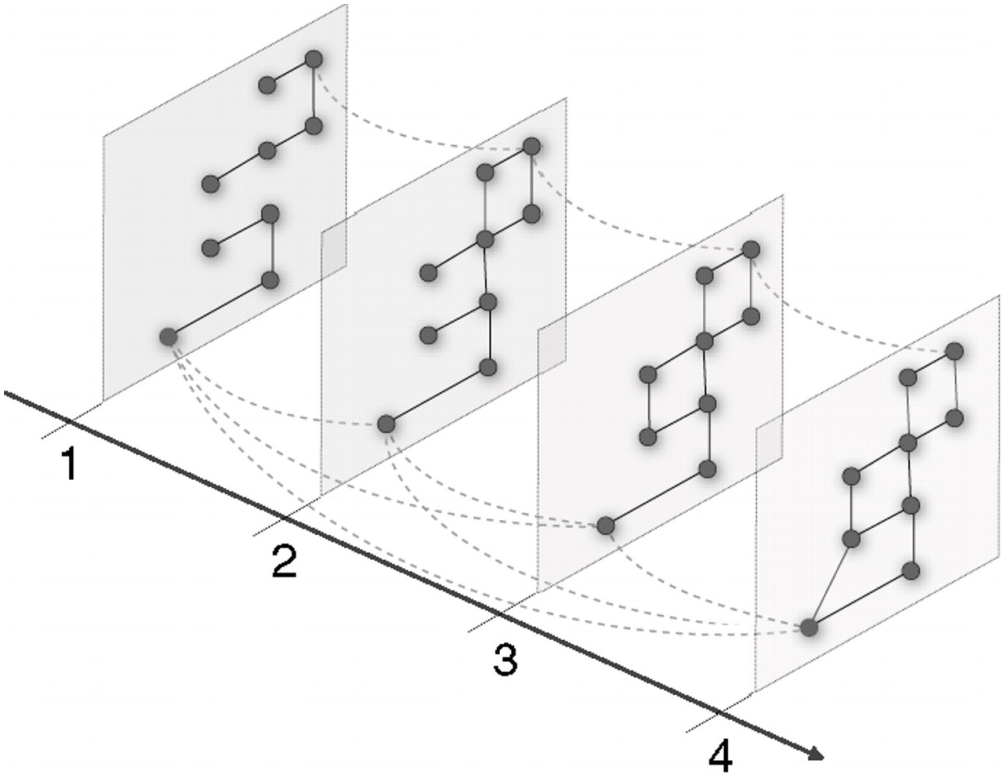
Smooth Back Reflector



Each simulation is performed with **at different ZnO**, used for the Transparent Conductive Oxide (TCO) layer, doping dosage (optimal, normal, low, not) and **at different Ag**, the Back Reflector used, roughness (smooth, rough) combination.

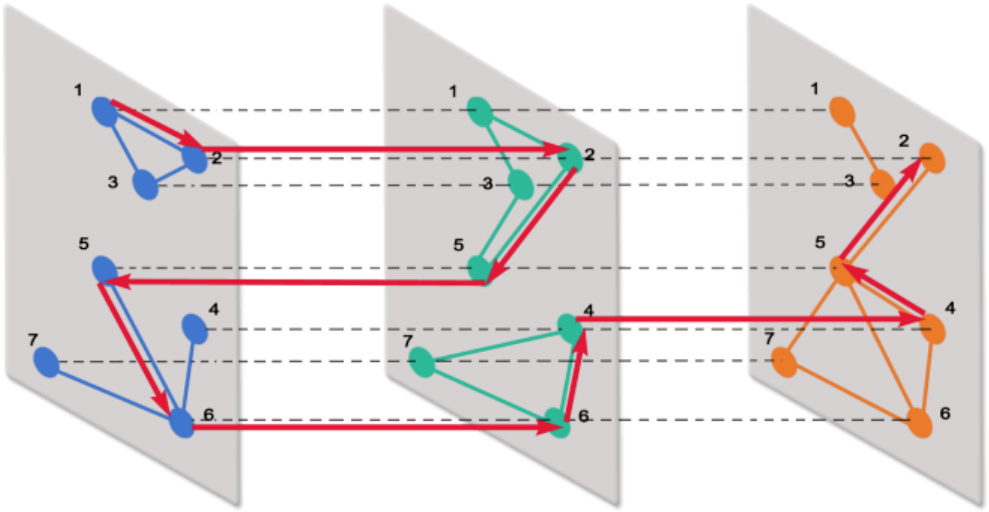
Airline Transportation Systems as Multiplex Networks

in collaboration with V. Latora, V. Nicosia, A. Santoro - QMUL



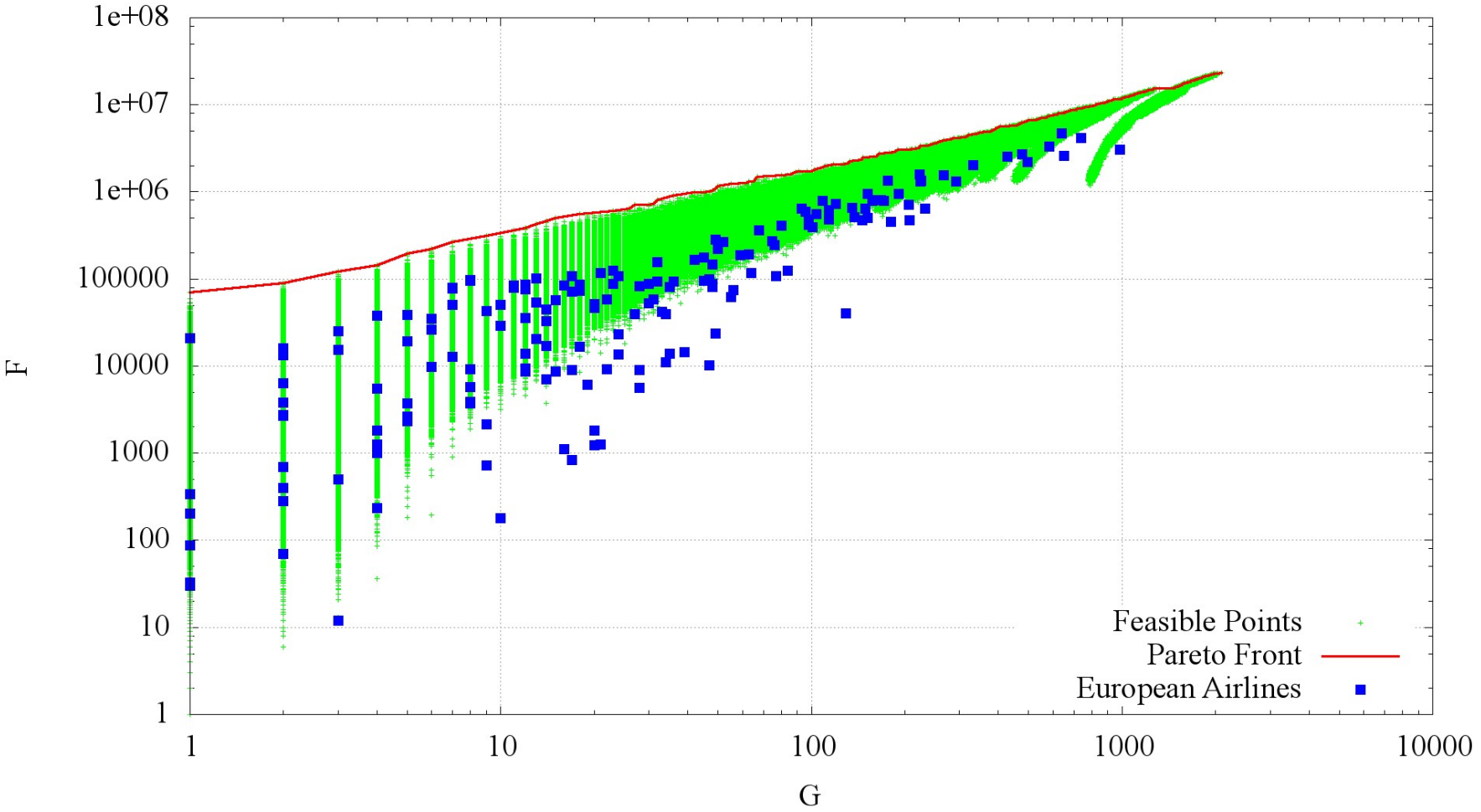
Airline Transportation Systems as Multiplex Networks

- We have a multiplex network with **N nodes** and **M layers**, where **each node stands for an airport** while **each layer represents a different airline**. This kind of network can be easily described as a set of adjacency matrix $\{A^{[1]}, A^{[2]}, \ldots, A^{[M]}\}$ in R^{NxNxM} in which
  - $a_{ij}^{[\alpha]} = 1$, if i and j are connected by a link (that represents the existence of a flight operated by the α-th airline) or
  - $a_{ij}^{[\alpha]} = 0$ otherwise.



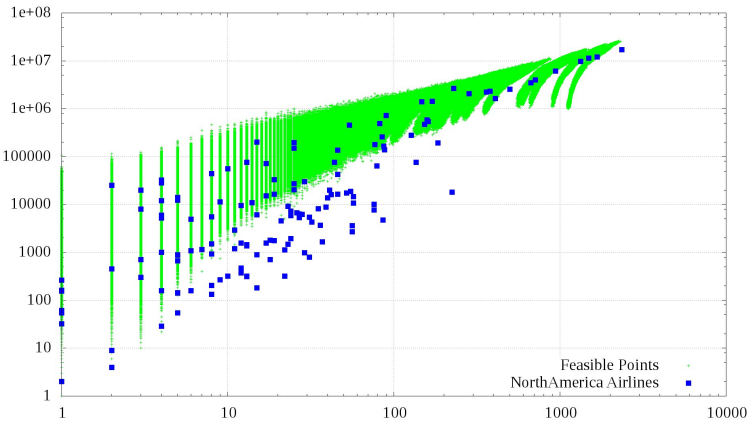
- **Max F** = **maximize** *the number of potential customers*
- **Min G** = **minimize** *the competition with other flight companies*, i.e. it should avoid to place a route between airport i and airport j if another company already provides a connection between those airports.

Pareto Front in the Model and in the Real **European Airlines**

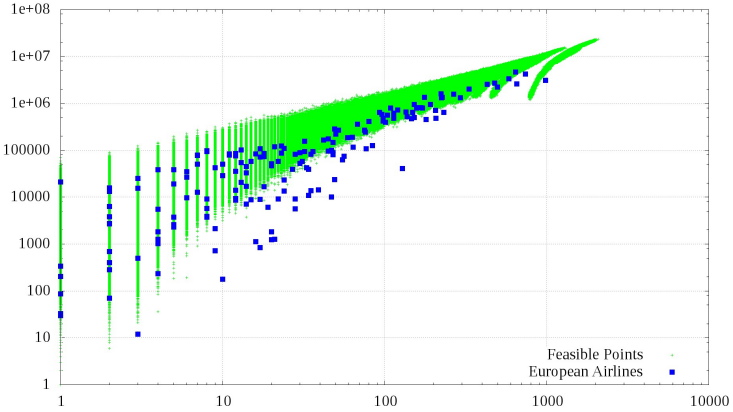


Maximize F the number of potential customers versus
Minimize G the competition with other flight companies

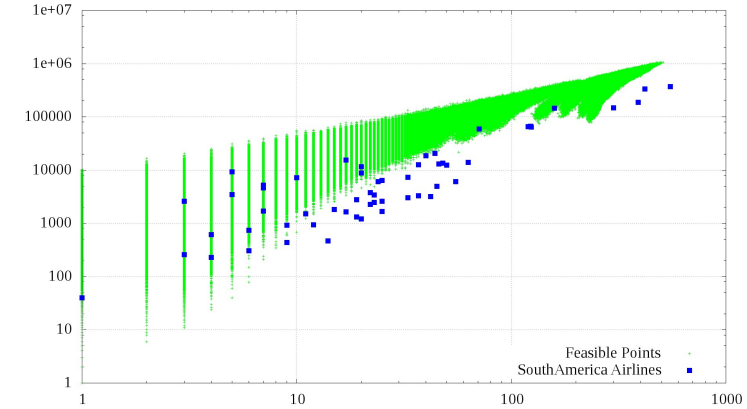
North America Airlines



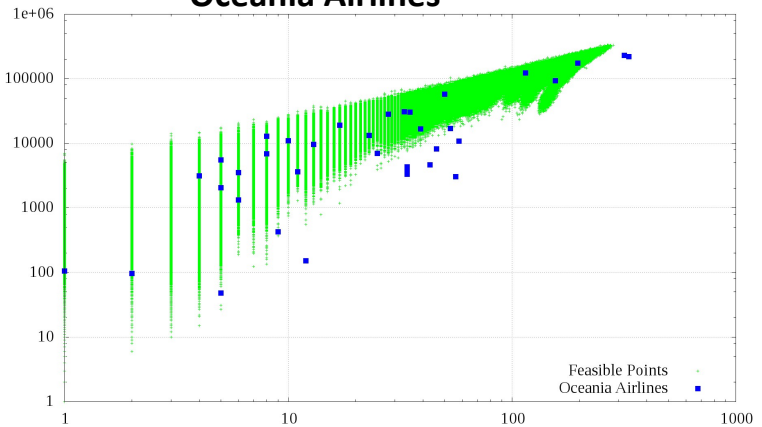
Europe Airlines



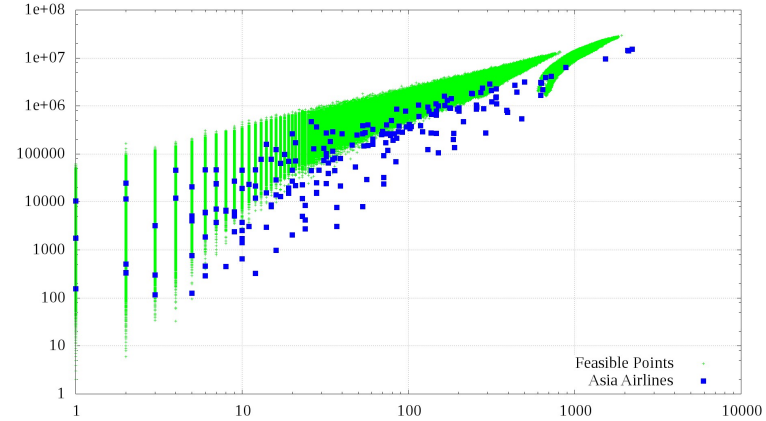
South America Airlines



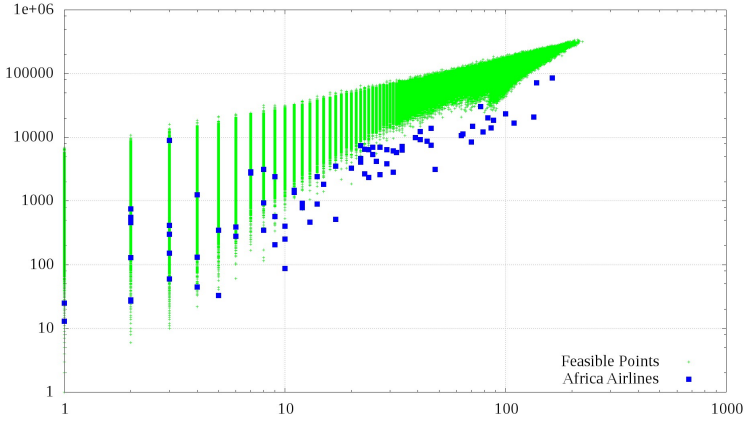
Oceania Airlines



Asia Airlines



Africa Airlines



Designing High Performance Semiconductors

- MESFETs
- MOSFETs
- Double Gate MOSFETs

Finding an optimal doping profile of a semiconductor device: the case of P-N silicon diode

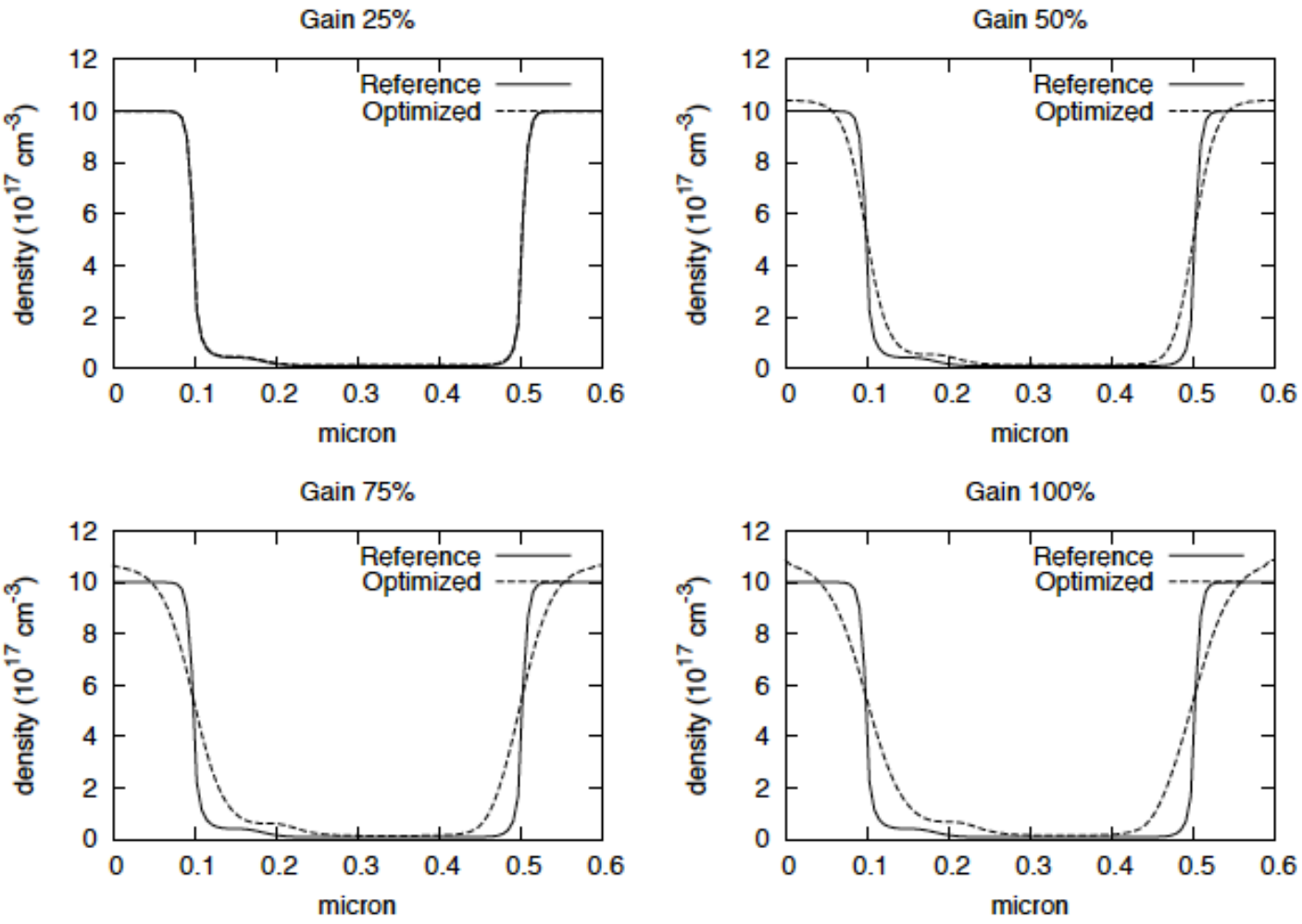
- The device: The n^+nn^+ diode is an unipolar model for the channel of a MOS transistor, where the current density is dominated by the electrons. The reference system is 0.6 μm long with a ***maximal doping concentration*** of $c_0 = 10^{18}$ cm^{-3} in the n^+ regions, $c_1 = 10^{16}$ cm^{-3} in the n region and an applied voltage of 1.0V; this device achieves a current of $J = 2.571 \text{ X } 10^4$ V/μm.

- Goal: to obtain an amplification of the output current; a gained output current Jg can be obtained by slightly changing the doping profile.

- Firstly, we adopt a drift-diffusion model for device simulation [Jungel et al 2001]; this class of models combines an accurate description of physical phenomenon with a low computational cost [Burger 2003].

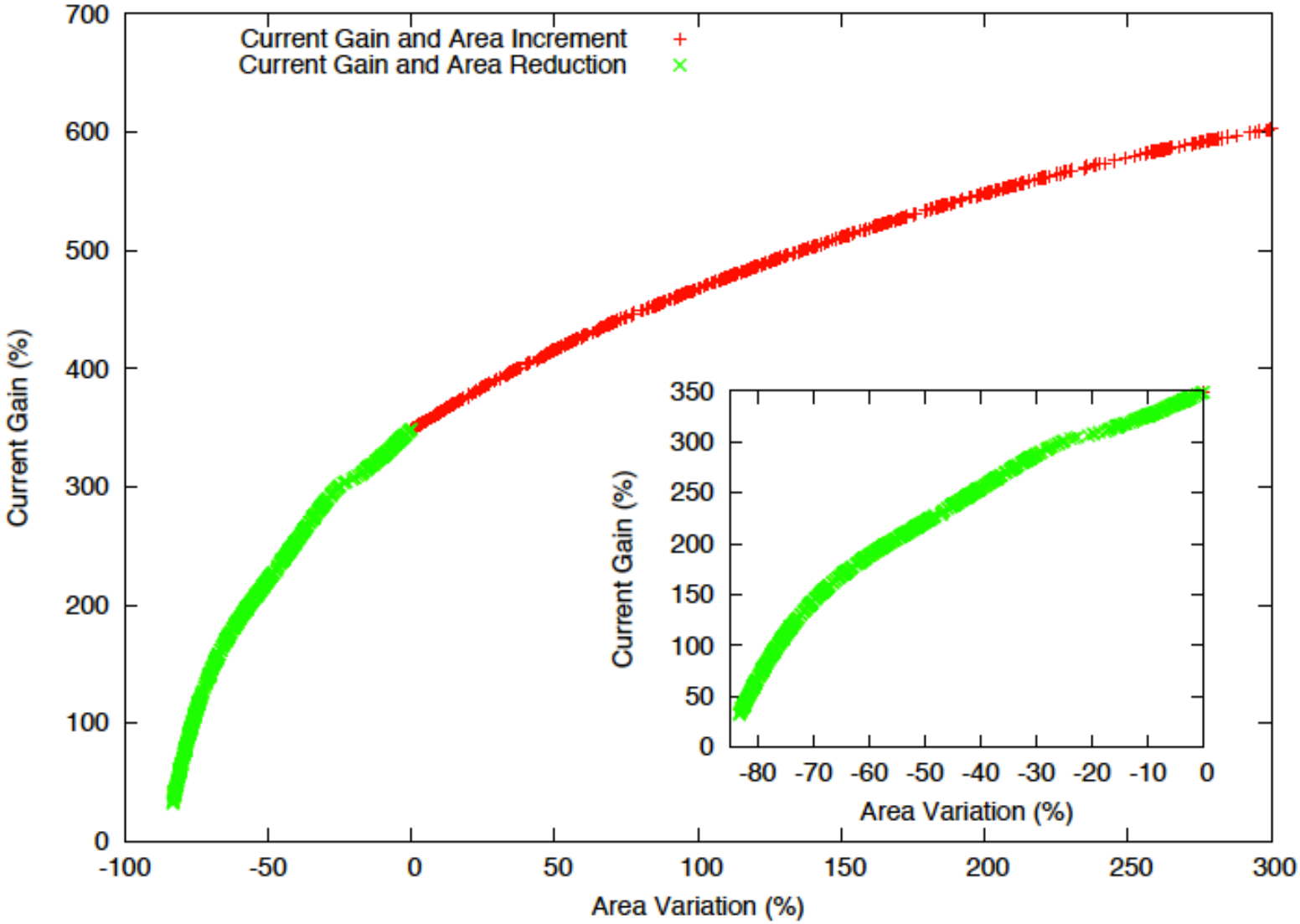
- Successively, the doping profile optimization problem has been tackled as a *constrained optimization problem*} (COP) [Biondi & Nicosia '06]; in particular,
  - we maximize the current gain (**objective function**);
  - by allowing at most θ deviation from the reference doping profile (**constraint**);
  - the **design parameters**: c_0, c_1 and s is a (hidden and given in this application) parameter that controls *the shape of the doping*.

Optimized n^+nn^+ silicon diode doping profiles



We perform a comparison of optimized doping profiles with the reference doping, at different current gain levels.

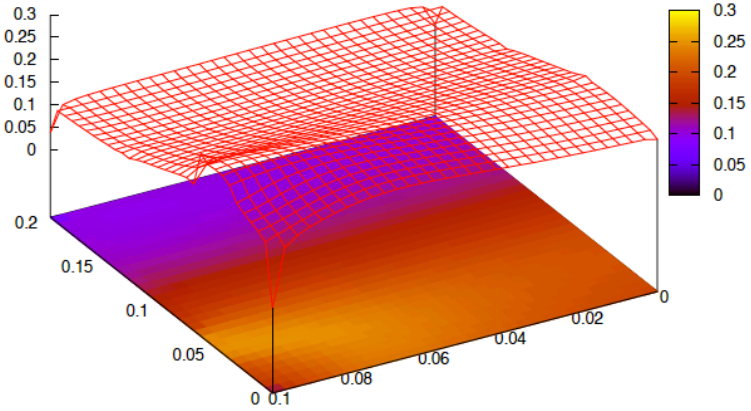
Pareto Front: Current Gain vs. Area



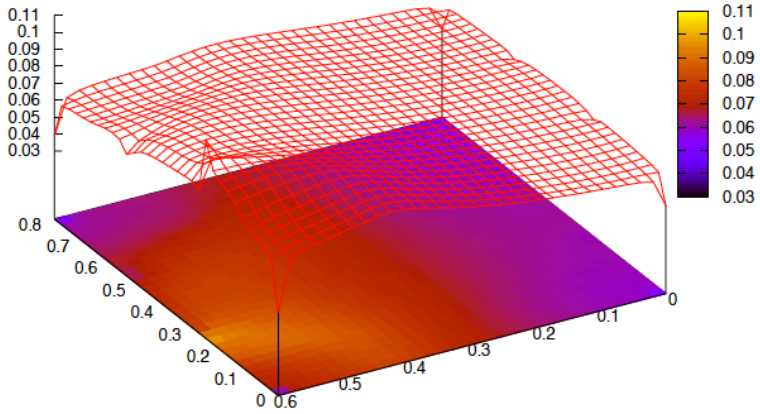
MESFET

Selection	Area (%)	Js (%)
Min Area	-83.33	31.94
Max Current	300.00	602.59
Closest to Ideal	35.31	397.86

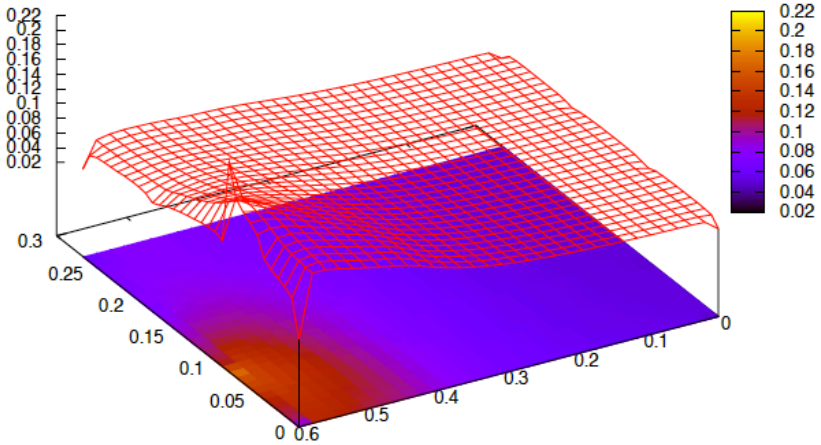
MESFET LOW-AREA - Energy (eV)



MESFET HIGH-CURR - Energy (eV)

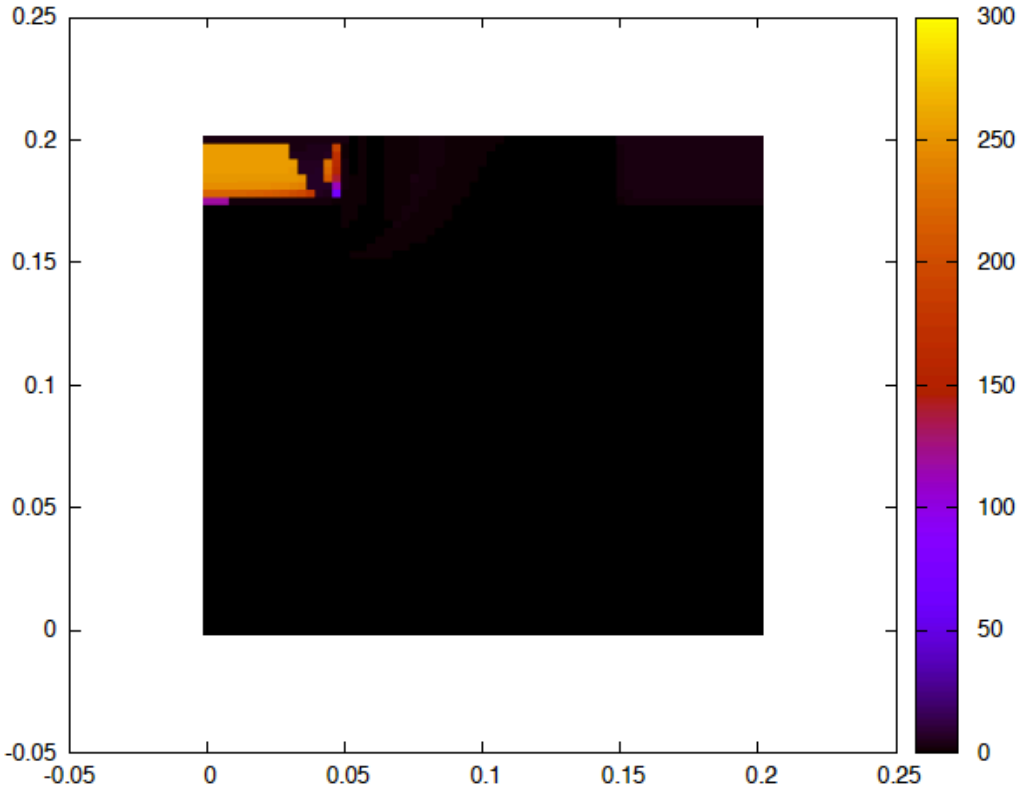


MESFET CTI - Energy (eV)



MOSFET

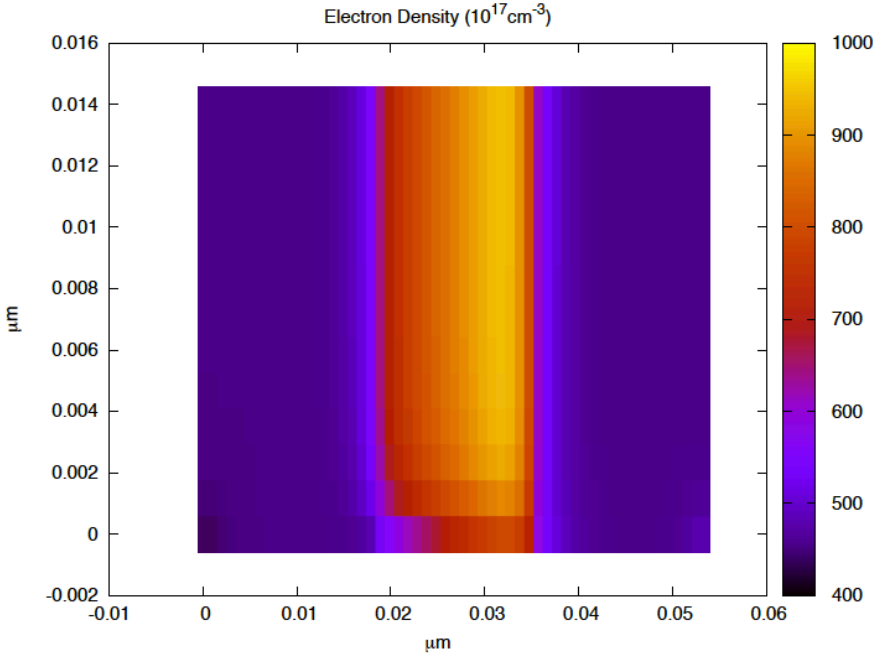
Selection	Area (%)	Js (%)
Min Area	-75.00	129.96
Max Current	-55.41	266.19
Closest to Ideal	-75.00	178.85

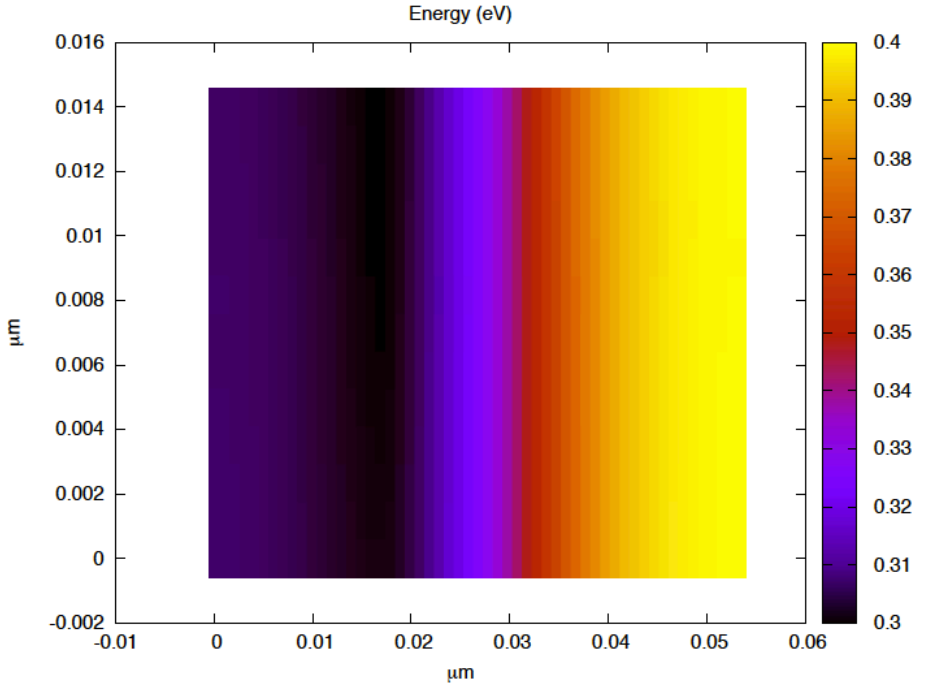


Energy distribution of the MOSFET device closest to ideal (Pareto optimality).

Double Gate MOSFET

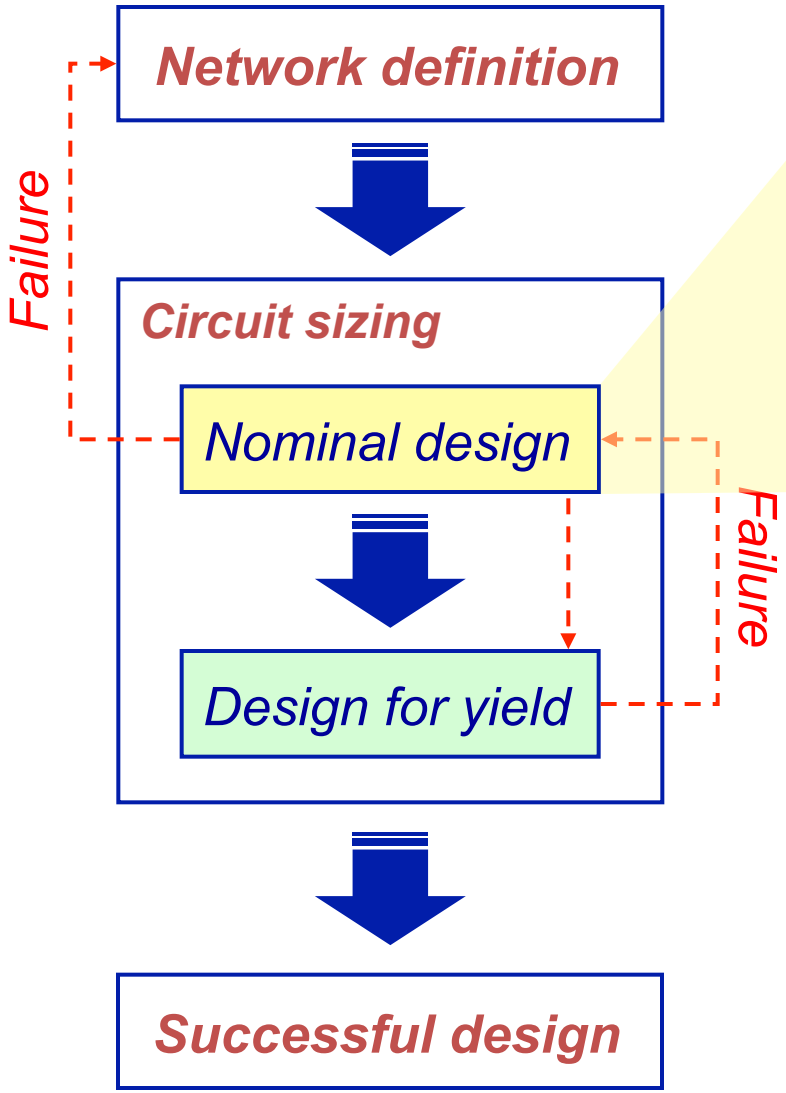
Selection	Area (%)	Js (%)
Min Area	-66.34	324.73
Max Current	-56.41	1256.39
Closest to Ideal	-64.31	1184.29





Designing High Performance Circuits

Nominal design



Define the set of parameters that optimizes circuit performance under fixed operating conditions:

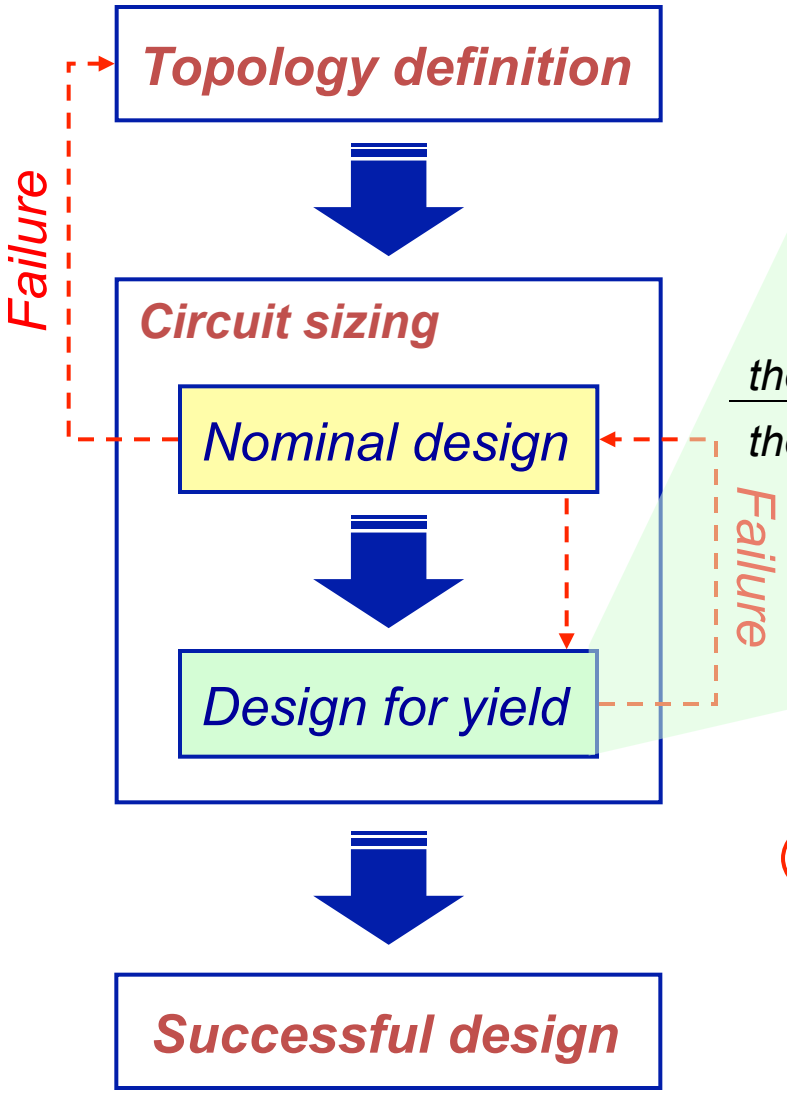
- *supply voltage*
- *ambient temperature*
- *fabrication process*

If **p** is a vector of parameters values on which the performance function f(**p**) depends, then the sizing problem is to find a set of values for **p** represented by **p0** such that **f(p0) < fg**

where **fg** is the set of *global specifications* for the circuit.

Failure mandates circuit topology to be modified

Design for yield



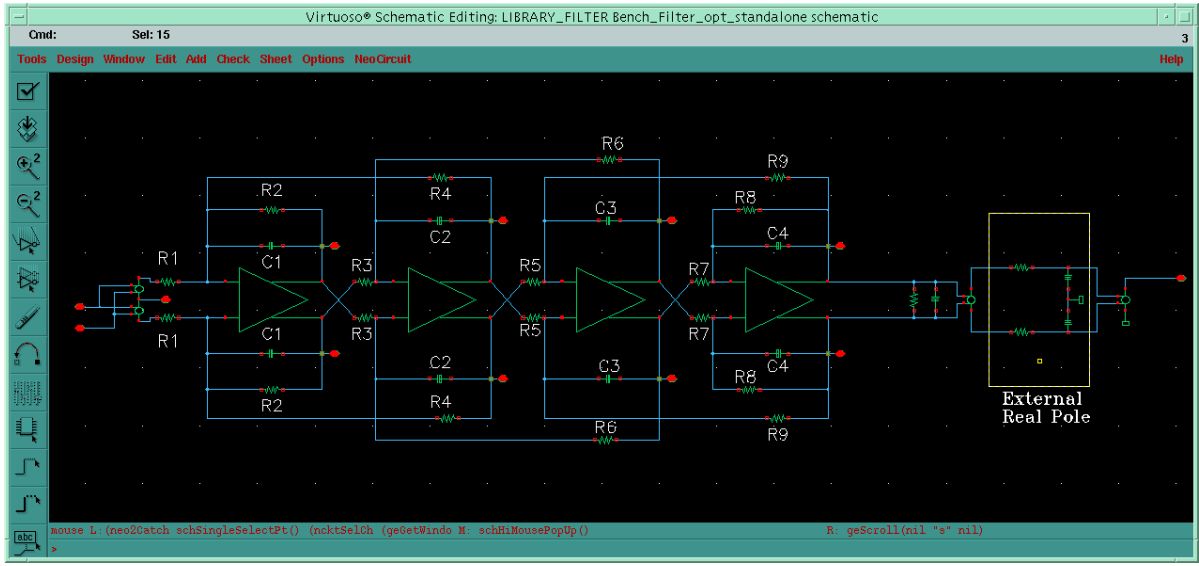
If the statistical variation of the parameters is to be considered,*e. g., process variations, then the circuit sizing problem is re-phrased as a yield improvement problem. The goal in such cases is to increase the percentage of circuit meeting the goal specifications (**fg**) for a particular set of input parameters (**p0**):*

$$\frac{\text{the number of chips that have acceptable performance}}{\text{the number of chips that have been manufactured}}$$

Failure requires changes to nominal design and iterations

Leapfrog Filter for W-LAN

1 run = 1 minute



Objectives and constraints (3+13)

- GroupDelayRipple@9.1MHz < 20 ns
- GroupDelayRipple@9.7MHz < 40 ns
- GroupDelaySlope@6MHz < 3 fs/Hz
- PassBandGain > -0.01 dB
- InputNoise < 44 nV/Hz$^{1/2}$ **(minimize)**
- PassBandRipple@9.1MHz < 0.8 dB
- PassBandRipple@9.7MHz < 1.8 dB
- StopBand@22.5MHz > 25 dB
- StopBand@34.2MHz > 56 dB
- DC current < 40 mA **(minimize)**
- OutputSwingOA1 < 2.8 V
- OutputSwingOA2 < 2.8 V
- OutputSwingOA3 < 2.8 V
- OutputSwingOA4 < 2.8 V
- InputResistance > 12.2 KOhm
- TotalArea < 18.000 um^2 **(minimize)**

Optimization variables (20)

- C = [1fF : 1fF :600fF]
- C_1 = [0.725n : 0.005n : 8.7n]
- L_2 = [18.625n : 0.01n : 223.5n]
- C_3 = [2.1n : 0.005n : 25.2n]
- L_4 = [11.875n : 0.01n : 142.5n]
- w_0 = [0.255 : 0.001 : 3.06]
- $m_{1,2,3,4}$ = [0.01 : 0.01 : 60]
- $k_{1,2,3,4}$ = [0.01 : 0.01 : 60]
- R_a = [1 : 0.005 : 12]
- wrp = [14.125MHz : 0.005MHz : 169.5MHz]
- $Vn_{1,2,3,4}$ = [5 : 0.05 : 60]

Leapfrog Filter: Comparisons

(M.J.D. Powell)

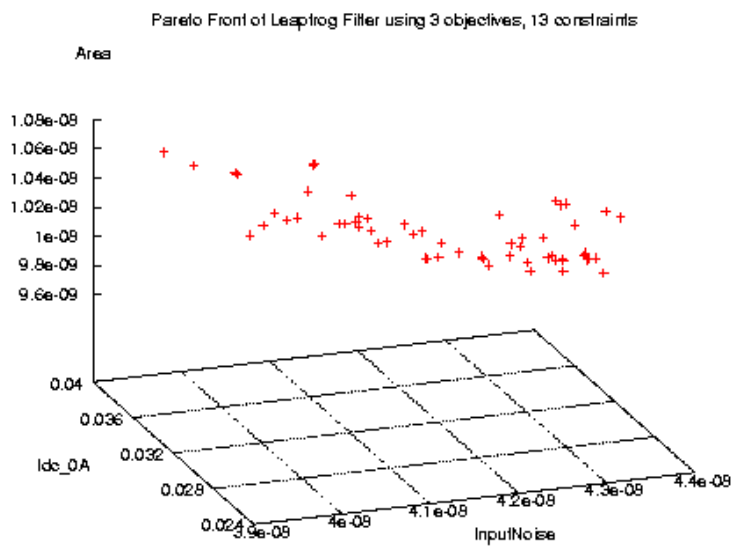
cadence

Performance Function	Constraint	NEWUOA [12]	DIRECT [11]	A-CRS [9]	Designer	A-NSGAII	optIA	Com. sol. [4]
DC gain	≥ -0.01 dB	−0.003	−0.003	−0.0026	−0.003	−0.0025	−0.003	−0.003
Pass-band ripple at 9.1MHz	$\leq 0.8\ dB_{PP}$	**0.95**	**0.86**	**0.85**	**0.806**	0.55	0.74	0.798
Pass-band ripple at 9.7MHz	$\leq 1.8\ dB_{PP}$	0.97	1.33	1.07	0.97	0.55	0.74	0.798
Stop-band at 22.5MHz	$\geq 25\ dB_C$	26.60	36.56	31.86	36.64	36.84	37.06	36.1
Stop-band at 34.2MHz	$\geq 56\ dB_C$	**46.11**	**55.08**	**51.64**	**55.92**	56.23	56.13	**54.9**
Group-delay ripple at 9.1MHz	≤ 20 ns	**24.52**	20.7	14.80	19.28	16.17	19.42	19.9
Group-delay ripple at 9.7MHz	≤ 40 ns	30.55	39.6	16.50	26.36	24.18	29.97	30.5
Group-delay slope at 6.0MHz	≤ 3 fs/Hz	2.63	1.28	**3.23**	2.10	2.36	2.48	1.4
Equivalent input resistance	$\geq 12.2\ k\Omega$	**11.82** k	31.54 k	13.42k	**12.18** k	12.24 k	13.56	12.8
Output dynamic of stage 1	≤ 2.8 V	**2.88**	2.23	2.72	**2.83**	2.76	2.52	2.79
Output dynamic of stage 2	≤ 2.8 V	2.68	1.36	2.20	2.22	2.43	2.76	2.59
Output dynamic of stage 3	≤ 2.8 V	**3.17**	2.60	**3.01**	1.71	2.70	2.66	2.79
Output dynamic of stage 4	≤ 2.8 V	2.13	1.32	1.43	1.53	1.73	1.42	1.25
Equivalent input noise	$\leq 44\ nV/Hz^{\frac{1}{2}}$	**47.73**	**139**	**47.41**	**46.44**	42.14	43.32	43.7
Dc current consumption	$\leq 40mA$	**40.64**	30	**42.17**	39.58	34.47	32.95	39
Silicon area	$\leq 18000\mu m^2$	15964	**29500**	14457	14037	14622	14106	12950
Global Error		85.95%	289.1%	43.9%	7.8%	0	0	1.96%
Yield		n.a.	n.a.	n.a.	n.a.	69.5 %	58.6%	n.a.

Other Comparisons:

- **Mesh Adaptive Direct Search** (MADS), MATLAB GADS package & C/C++ code (C. Audet & J. Dennis)
- **Generalized Pattern Search Algorithms** GPS (C. Audet & J. Dennis)

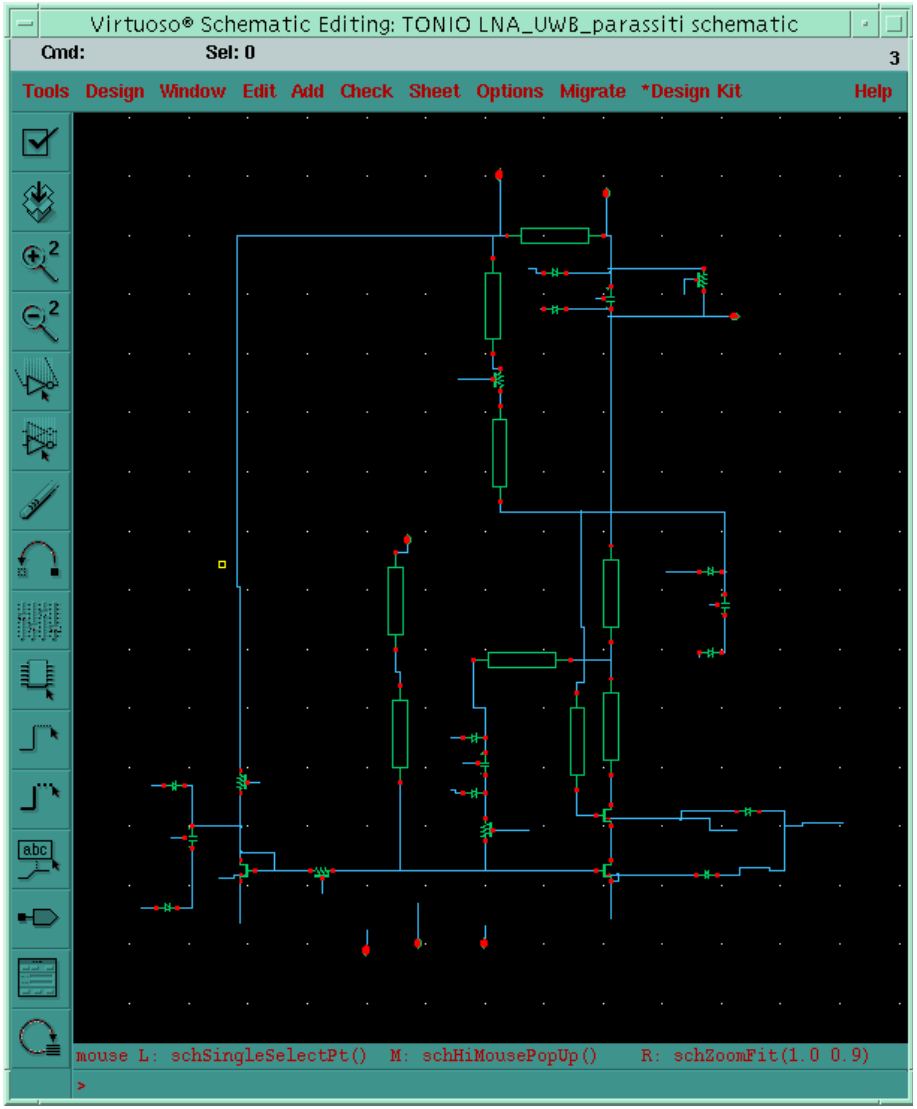
clA



Leapfrog Filter: Cadence vs. clA

	Objectives + constraints	Error	First feasible point/yield	Pareto/feasible/total run points	Yield*
ADE GLX (Global sizing)	3+13	**7%**	n.a.	-/-/23000	n.a.
A-NSGAII	3+13	0%	22500/49%	2/2/23000	45.2%
A-NSGAII	3+13	0%	22500/49%	**75**/**199**/36000	**69.4%**
ADE GLX (Global sizing)	1+15	**1.96%**	n.a.	-/-/36000	n.a.
clA	1+15	0%	25880/2%	**16**/**70**/36000	**58.6% (best nominal)**

Low Noise Amplifier for W-LAN



Optimization variables (24)

- L_{Gext} = [1nH : 0.1nH : 3.5nH]
- ***M₀, M₇, M₃₉*** { W_x = [0.12um : 0.01um : 50um]; L_x = [0.1um : 0.01um : 1um]; M_x = [1 : 1 : 200]; NF_x = [1 : 1 : 20] }
- ***R₁, R₃, R₄*** { W_{Rx} = [0.44um : 0.01um : 21.8um]; L_{Rx} = [2um : 0.01um : 400um]; M_{Rx} = [1 : 1 : 10] }
- W_{C0} = [3.5um : 0.01um : 30um]
- L_{C0} = [3.5um : 0.01um : 30um]

Objectives and constraints (2+4)

- S_{11} < -10 dB
- NF < 1.3 dB **(minimize)**
- AV_{4GHz} > 16.5 dB
- AV_{RIPPLE} < 3.0 dB
- IC_{LNA} < 4.4 mA **(minimize)**
- IC_{MIRROR} < 0.8 mA

1 run = 20 seconds

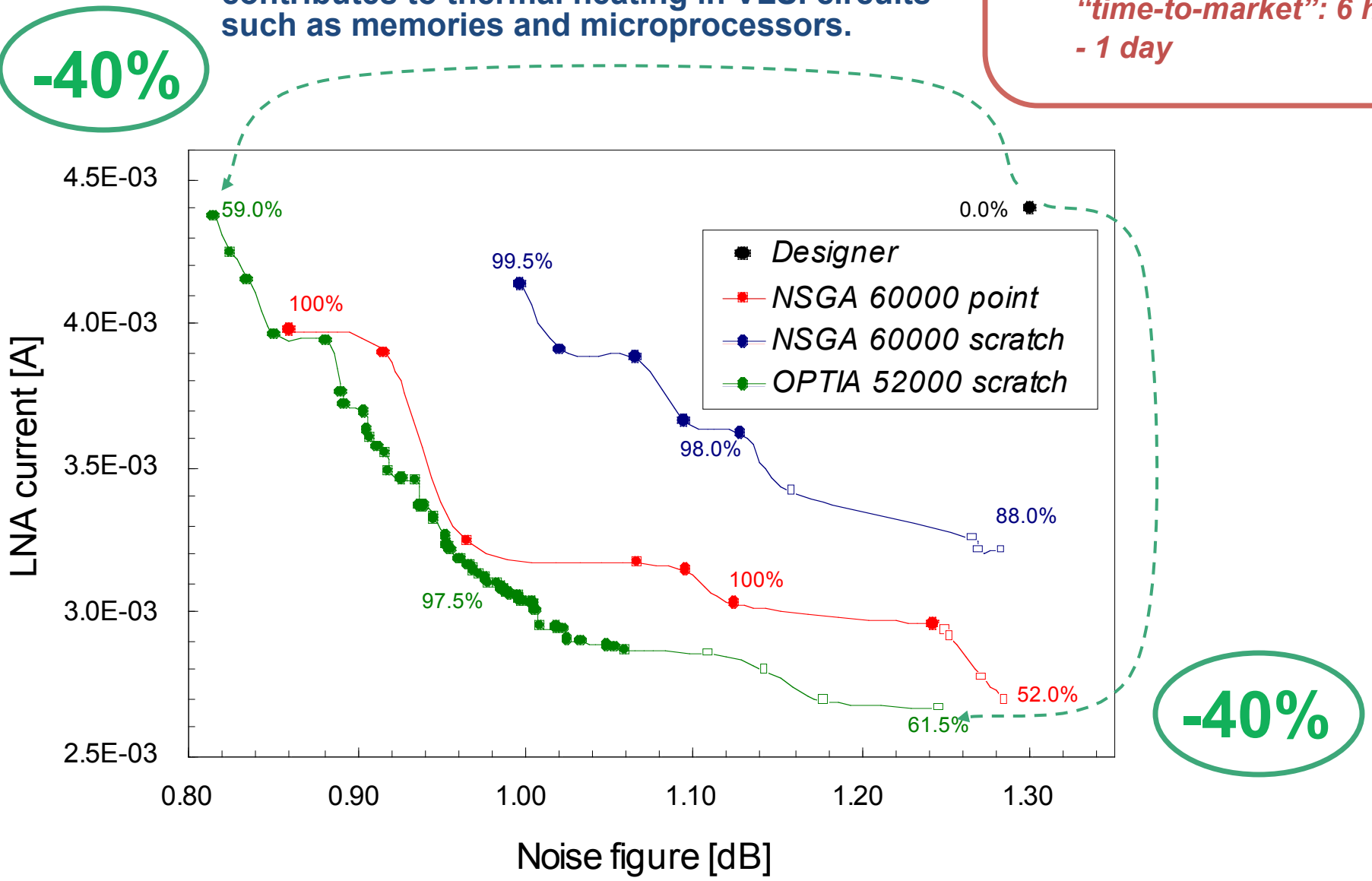
LNA 3-5 GHz results

Initialization method	Algorithm	First Feasible Circuit @ run	60000 Runs Feasible p./non dominated p.	Yield
Random	A-NSGAII	16722	100 / 9	100%
	cIA	*3963*	4747 / 55 All under-threshold circuits!	100%
From designer's point (error 8.46%)	A-NSGAII	1596	212 / 11	100%
	cIA	7582	2640 /36	100%

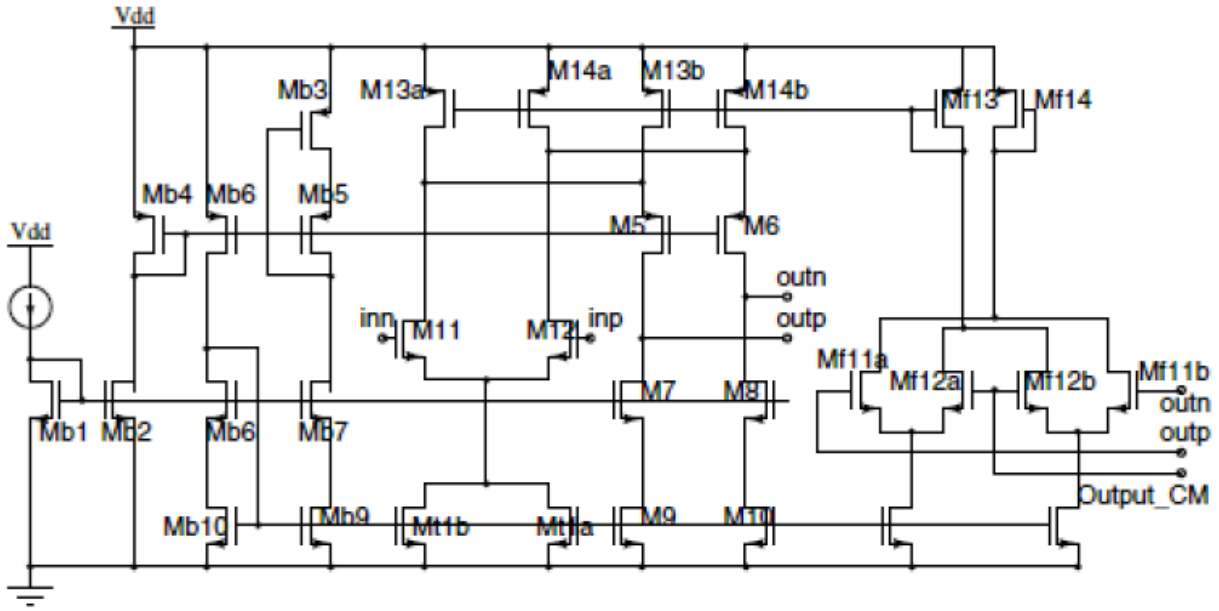
All under-threshold operation of MOS transistors!

Power consumption determines battery life in portable applications such as mobile phones or contributes to thermal heating in VLSI circuits such as memories and microprocessors.

1. *Best Performance*
2. *Maximum Yield: 100%*
3. *Under-threshold*
4. *Minimization of "time-to-market": 6 hours - 1 day*



Corner Analysis for Folded Cascode op-amp 1/2



Specification	Goal
DC gain(dB)	≥ 50
phase margin(°)	≥ 60
Gm(mS)	$\in [0.5\ ,\ 0.6\ ]$
CMFB DC gain(dB)	≥ 60
CMFB unity-gain bandwidth(MHz) (ω_{uCMFB})	≥ 5.9

Corner Analysis for Folded Cascode op-amp 2/2

Specification	Goal	YdIRCO[12]	NSGA-II[22]	**cIA**
yield(%)	maximize	94	95	**98**
area(μm^2)	minimize	668	631	**582**
power(mW)	minimize	0.33	0.32	**0.30**

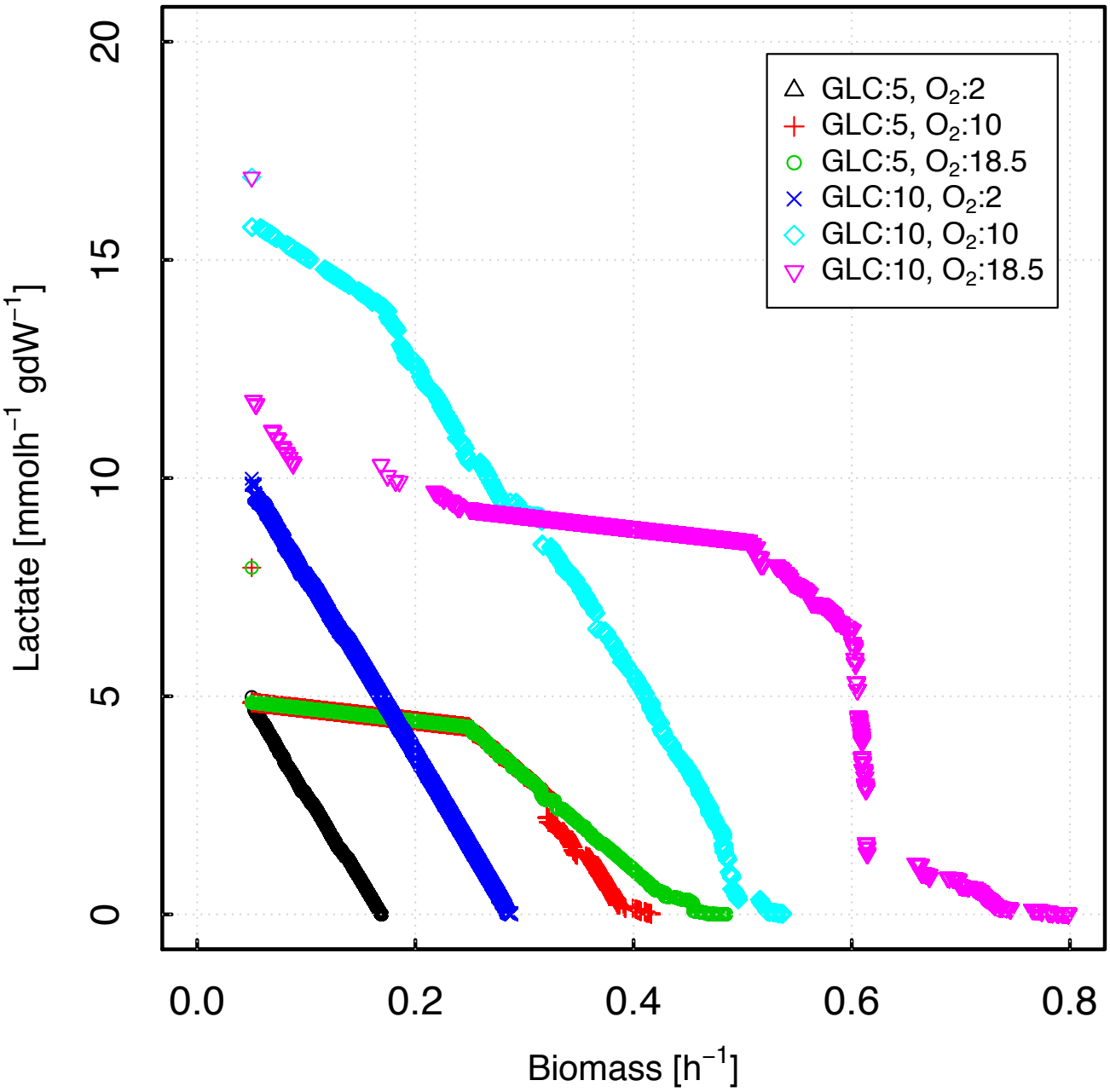
Robust Folded Cascode op-amp Design at three different temperatures.

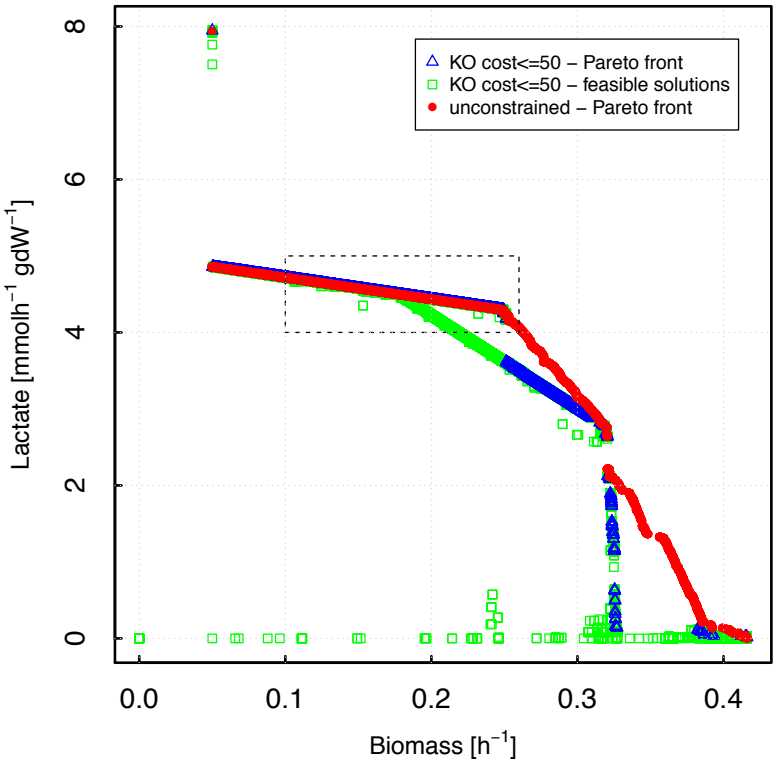
Specification	Goal	NSGA-II[22]	cIA
DC Gain @ Corner 1 (303 K)	> 90	99	**112.215**
DC Gain @ Corner 2 (333 K)	> 90	95	**118.193**
DC Gain @ Corner 3 (423 K)	> 90	106	**120.092**
THDse @ Corner 1 (303 K)	< -25	-37.219	**-48.754**
THDse @ Corner 2 (333 K)	< -25	-31.763	**-44.271**
THDse @ Corner 3 (423 K)	< -25	-29.165	**-34.194**
THDd @ Corner 1 (303 K)	< -65	-75.179	**-79.184**
THDd @ Corner 2 (333 K)	< -65	-73.491	**-81.053**
THDd @ Corner 3 (423 K)	< -65	-72.236	**-86.184**
GBW @ Corner 1 (303 K)	$> 900M$	450M	**1101M**
GBW @ Corner 2 (333 K)	$> 900M$	471M	**1123M**
GBW @ Corner 3 (423 K)	$> 900M$	482M	**1201M**
TsL @ Corner 1 (303 K)	Min ($< 6n$)	4.9121	**1.196n**
TsL @ Corner 2 (333 K)	Min ($< 6n$)	3.761	**1.415n**
TsL @ Corner 3 (423 K)	Min ($< 6n$)	2.917	**1.921n**
TsH @ Corner 1 (303 K)	Min ($< 6n$)	3.492	**1.596n**
TsH @ Corner 2 (333 K)	Min ($< 6n$)	3.287	**1.717n**
TsH @ Corner 3 (423 K)	Min ($< 6n$)	2.957	**1.898n**

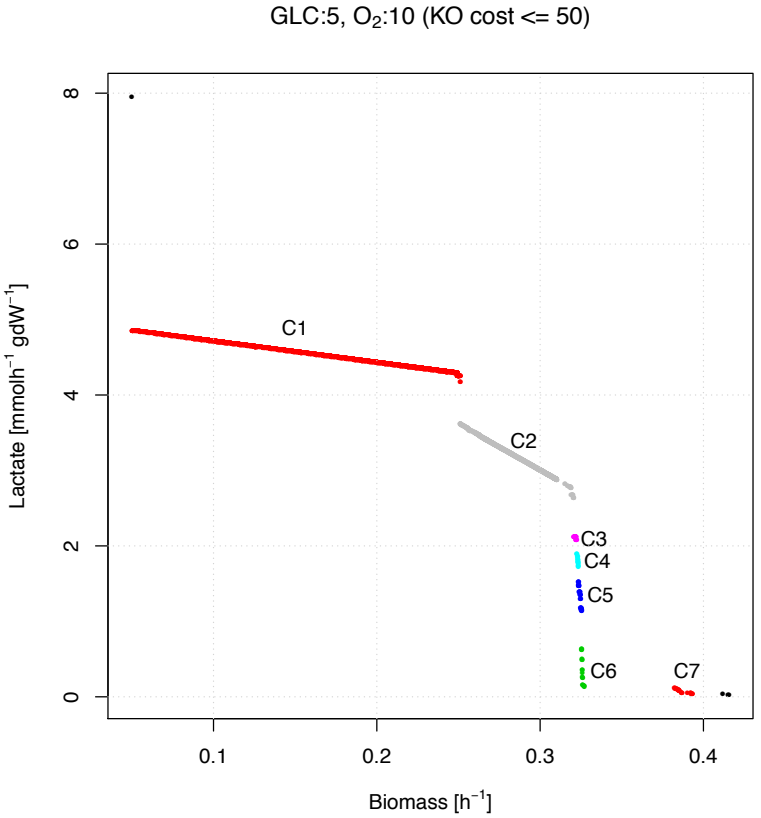
Designing BioPlastic: Polylactic acid from Lactate acid production in S. cerevisiae

Lactate Production in S. cerevisiae by Redesigning the Metabolic Networks

- Organism: Saccharomyces cerevisiae S288c
- Model: iMM904
- Genome: PRJNA128
- Metabolites: 1226
- Reactions: 1577
- Genes: 905
- Database: http://bigg.ucsd.edu/models/iMM904/
- Publication PMID: 19321003
- Mo ML, Palsson BO, Herrgård MJ., Connecting extracellular metabolomic measurements to intracellular flux states in yeast. BMC Syst Biol. 2009 Mar 25;3:37. doi: 10.1186/1752-0509-3-37.

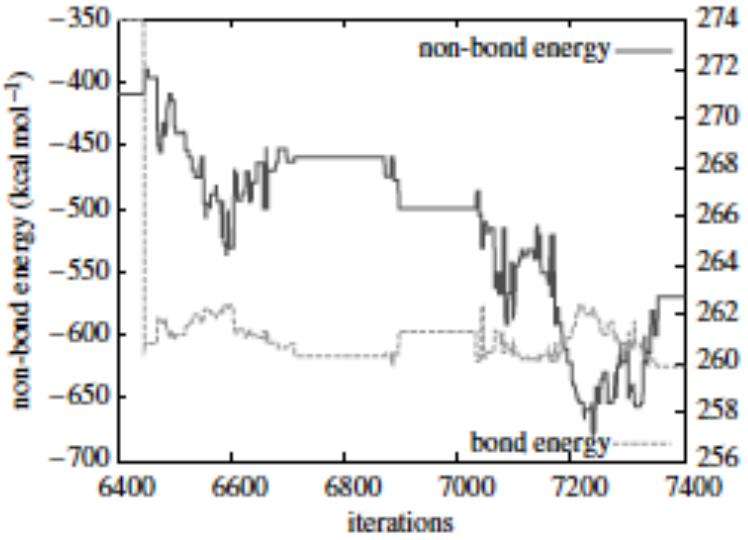

GLC:5, O2:2
GLC:5, O2:10
GLC:5, O2:18.5
GLC:10, O2:2
GLC:10, O2:10
GLC:10, O2:18.5
Lactate [mmolh−1 gdW−1]
Biomass [h−1]
0
5
10
15
20
0.0
0.2
0.4
0.6
0.8


KO cost<=50 – Pareto front
KO cost<=50 – feasible solutions
unconstrained – Pareto front
Lactate [mmolh^{-1} gdW^{-1}]
Biomass [h^{-1}]
0
2
4
6
8
0.0
0.1
0.2
0.3
0.4


GLC:5, O_2:10 (KO cost <= 50)
Lactate [mmolh^{-1} gdW^{-1}]
Biomass [h^{-1}]
C1
C2
C3
C4
C5
C6
C7
0
2
4
6
8
0.1
0.2
0.3
0.4

Protein Structure Prediction:
Pareto Fronts in Networks of Amino Acids

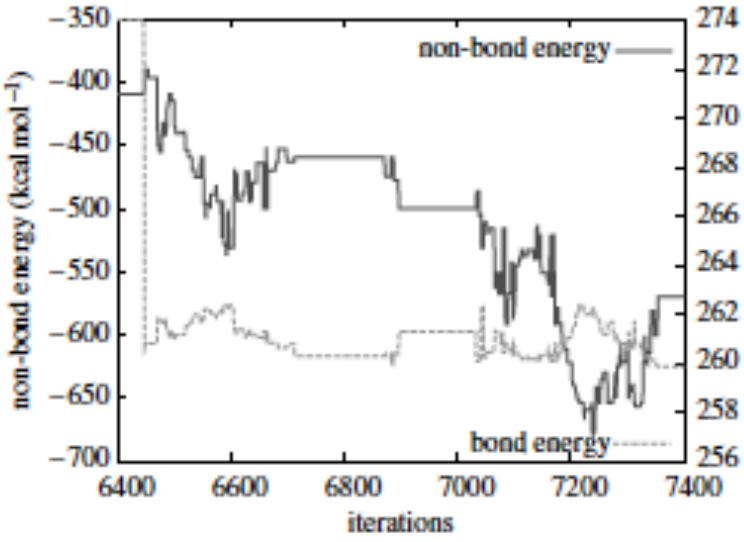
$$E_{\text{charmm}} = \underbrace{\sum_{\text{bonds}} k_b(b-b_0)^2}_{E_1} + \underbrace{\sum_{\text{UB}} k_{\text{UB}}(S-S_0)^2}_{E_2}$$
$$+ \underbrace{\sum_{\text{angles}} k_\theta(\theta-\theta_0)^2}_{E_3} + \underbrace{\sum_{\text{torsions}} k_\chi[1+\cos(n\chi-\delta)]}_{E_4}$$
$$+ \underbrace{\sum_{\text{impropers}} k_{\text{imp}}(\phi-\phi_0)^2}_{E_5}$$
$$+ \underbrace{\sum_{\text{non-bond}} \varepsilon_{ij}\left[\left(\frac{R\min_{ij}}{r_{ij}}\right)^{12} - \left(\frac{R\min_{ij}}{r_{ij}}\right)^{6}\right]}_{E_6}$$
$$+ \underbrace{\frac{q_i q_j}{\varepsilon r_{ij}}}_{E_7}, \qquad (3.2)$$

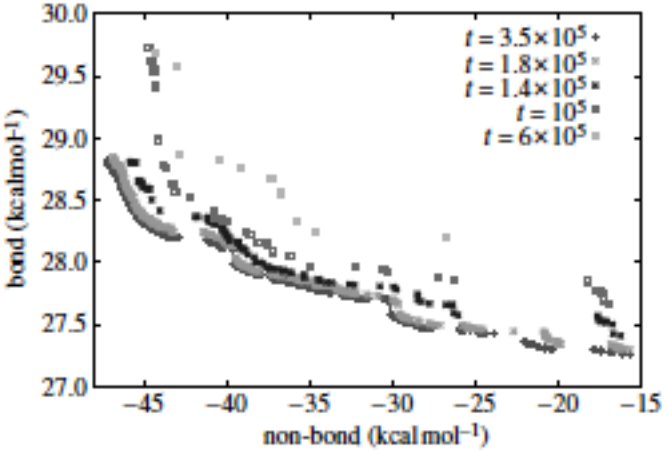


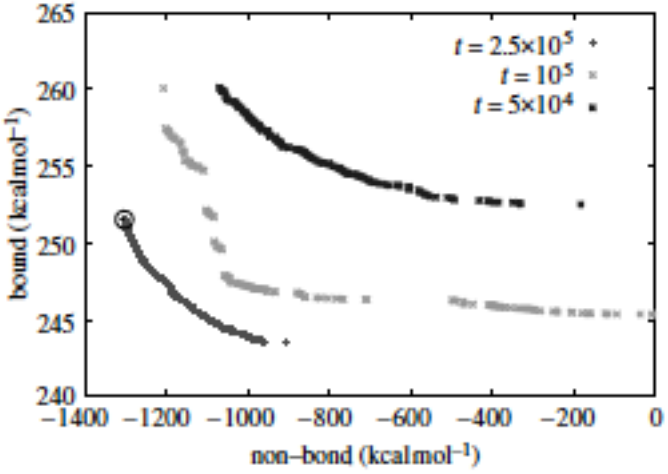
The Nobel Prize in Chemistry 2013

Martin Karplus, Michael Levitt, Arieh Warshel

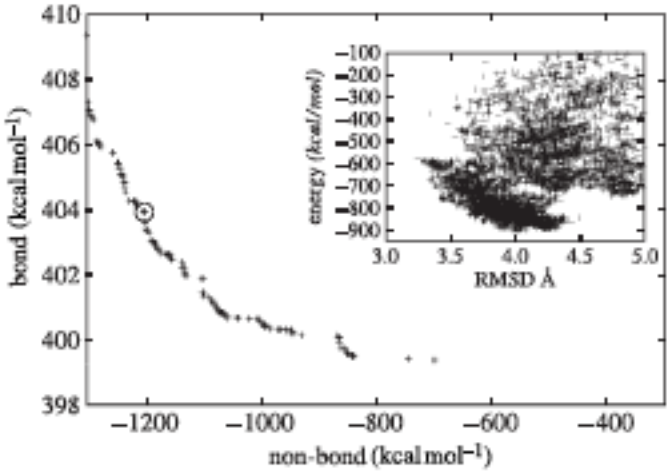
$$
\begin{aligned}
E_{\text{charmm}} = & \underbrace{\sum_{\text{bonds}} k_b (b - b_0)^2}_{E_1} + \underbrace{\sum_{\text{UB}} k_{\text{UB}} (S - S_0)^2}_{E_2} \\
& + \underbrace{\sum_{\text{angles}} k_\theta (\theta - \theta_0)^2}_{E_3} + \underbrace{\sum_{\text{torsions}} k_\chi [1 + \cos(n\chi - \delta)]}_{E_4} \\
& + \underbrace{\sum_{\text{impropers}} k_{\text{imp}} (\phi - \phi_0)^2}_{E_5} \\
& + \underbrace{\sum_{\text{non-bond}} \varepsilon_{ij} \left[\left(\frac{R \min_{ij}}{r_{ij}} \right)^{12} - \left(\frac{R \min_{ij}}{r_{ij}} \right)^{6} \right]}_{E_6} \\
& + \underbrace{\frac{q_i q_j}{\varepsilon r_{ij}}}_{E_7}, \qquad (3.2)
\end{aligned}
$$

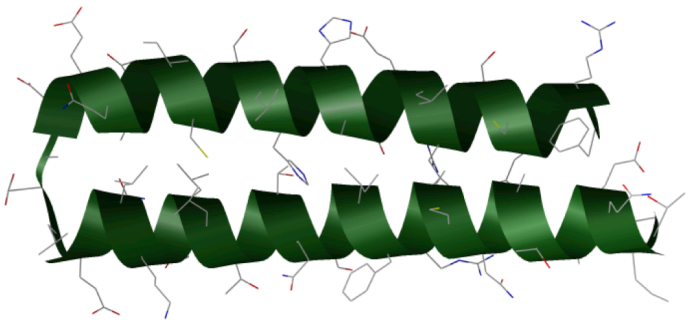


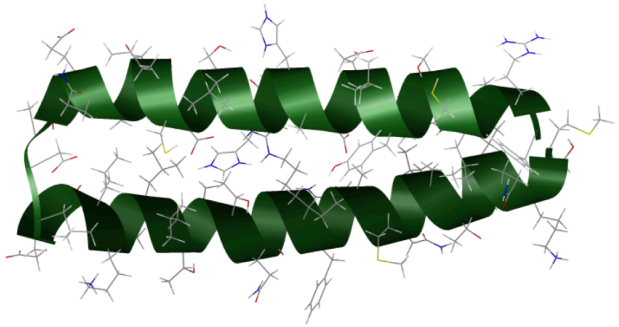




1ROP Protein







Min energy: -902.36 kcal/mol, RMSD=3.5 Angstrom, DME=1.62 Angstrom

References on Electronic Circuits and Systems

- "**Multi-Objective Optimization and Analysis for the Design Space Exploration of Analog Circuits and Solar Cells**", A. Patanè, A. Santoro, P. Conca, G. Carapezza, A. La Magna, V. Romano, and Giuseppe Nicosia, *Engineering Applications of Artificial Intelligence*, 2016 (to appear). DOI: 10.1016/j.engappai.2016.08.010

- "**Semiconductor Device Design using Bimads Algorithm**", G. Stracquadanio, V. Romano, Giuseppe Nicosia, *Journal of Computational Physics*, 242:304-320, 2013.
- "**Clonal Selection - An Immunological Algorithm for Global Optimization over Continuous Spaces**", M. Pavone, G. Narzisi, Giuseppe Nicosia, *Journal of Global Optimization*, 53(4):769-808, 2012.

- "**An Evolutionary Algorithm-Based Approach to Robust Analog Circuit Design using Constrained Multi-Objective Optimization**", Giuseppe Nicosia, S. Rinaudo, E. Sciacca, *Knowledge-Based Systems J*., 21(3):175-183, 2008.

References on Biological Circuits and Systems

- "**Pareto Optimal Design for Synthetic Biology**", A. Patane', A. Santoro, J. Costanza, Giuseppe Nicosia, *IEEE Transactions on Biomedical Circuits and Systems*, 9(4): 555-571, 2015. DOI:10.1109/TBCAS.2015.2467214

- "**Inferring Pathological States in Cortical Neuron Microcircuits**", J. Rydzewski, W. Nowak, Giuseppe Nicosia, *Journal of Theoretical Biology*, 386:34-43, 2015. DOI: 10.1016/j.jtbi.2015.09.004

- "**Robust Design of Microbial Strains**", J. Costanza, G. Carapezza, C. Angione, P. Lio', Giuseppe Nicosia, *Bioinformatics - Oxford Journal*, 28(23):3097-3104, 2012.

References on Physics-inspired computation: Satisfiability Problem, Phase Transition, Bose-Einstein Condensation & Maxwell-Boltzmann Distribution

- "**Satisfiability by Maxwell-Boltzmann and Bose-Einstein Statistical Distributions**", C. Angione, A. Occhipinti, Giuseppe Nicosia, *ACM Journal of Experimental Algorithmics*, Volume 19, October 2014, Article No. 1.4, doi:10.1145/2629498.

- "**Bose-Einstein Condensation in Satisfiability Problems**", C. Angione, A. Occhipinti, G. Stracquadanio, Giuseppe Nicosia, *European Journal of Operational Research*, 227:44-54, 2013.

Conclusions

Pareto Optimality as Pareto Law

Transistors/Amino Acids/Si atoms networks as MOO

Pareto Optimality for Prescriptive Big Data Analytics

Acknowledgments

- **Jole Costanza, Post-doc IIT Milan**
- **Piero Conca, Post-Doc**
- **Andrea Patanè, Oxford University**
- **Andrea Santoro, Queen Mary University London**
- **Giovanni Carapezza, Post-Doc**
- **Giorgio Jansen, PhD Student**

- **IBM** – Rome, Italy
- **MAXIM Semiconductors** – Italy
- **DIALOG Semiconductors** – Germany & Austria.

Thanks!