How to become Physically Net Zero by 2030

Bruno Michel

Mgr. Smart System Integration
IEEE Fellow, Member IBM Academy of Technology, and US National Academy of Engineering
Agenda

- Split Sustainability Problem to Solve ... Efficiency is Key!
- Density and Technology Generation linked to Efficiency
- Stairway to paradise for HPC and Cloud
- From Cloud to Edge ... Proximity Improves efficiency
- Stop Red AI (Energy waste) ... Start Green AI
- Carbon capture removes residual emissions ... reuse of low thermal resistance technologies
IBM Commitment: Physical Net Zero GHG Emissions by 2030

IBM is updating its goals to achieve net zero greenhouse gas emissions by 2030 comprehensive – global

- Reduce greenhouse gas emissions 65% by 2025 against 2010.
- Procure 75% electricity from renewables by 2025, and 90% by 2030.
- Use technologies, such as carbon capture (in or by 2030) to remove emissions in an amount which equals or exceeds residual emissions.

Why companies claim net zero status as of now? ➔ Two different yardsticks!

Financial versus Physical Net Zero

Financial Net Zero is 10-30x cheaper than Physical Net Zero

RECs and Financial Offsets (planting trees in 3rd world) are “innovation killers”

IBM first to claim reaching PHYSICAL Net Zero by 2030

This is ~100x more difficult and ~100x more impactful!
What is the Problem?

Worldwide carbon emissions 2019 = 37 Gt, ICT ~1 Gt = ~3% 2..4%
Fast upwards trend for ICT power


Strongly growing carbon footprint of Cloud datacenters ...
Data transport energy is a large fraction
AI energy doubles every 3-4 months
Split the Problem to Solve it Better

$$\text{CFP} = \text{E}_{\text{IT}} \times \text{R}_{\text{CPF}}$$

$$\text{R}_{\text{CPF}} = \{(\text{E}_{\text{IT}} + \text{E}_{\text{DC}}) \times \text{CI} - \alpha \times \text{E}_{\text{reuse}}\}/\text{E}_{\text{IT}} \quad \text{R}_{\text{CPF}}: \quad >2.0 \ldots -1.0$$

$$\text{PUE} = (\text{E}_{\text{IT}} + \text{E}_{\text{DC}})/\text{E}_{\text{IT}} \quad \text{PUE}: \quad >2.0 \ldots 1.05$$

$$\text{ERE} = (\text{E}_{\text{IT}} + \text{E}_{\text{DC}} - \text{E}_{\text{reuse}})/\text{E}_{\text{IT}} \quad \text{ERE}: \quad >2.0 \ldots 0.0$$

$$\alpha \times \text{E}_{\text{reuse}} \text{ (physical offset of fossil)} \quad \alpha: \quad 1..0 \quad \text{E}_{\text{reuse}}: \quad 0 \ldots \text{E}_{\text{IT}} + \text{E}_{\text{DC}}$$

ERE is an extension of the more well known PUE that includes physical offset.

Carbon Intensity (CI) is >1 kg/kWh for coal power, <0.1 kg/kWh for solar and wind power (no RECs).

E_{IT} is composed of communication energy E_{comm} and computation energy E_{comp}.

E_{comp} in a computer is composed of energy for data movement E_{move} and processing E_{proc}.
Datacenter Efficiency

Efficiency is 0.000’004%
Volume-fraction for computing is <1 ppm (part per million)

A computer is an inefficient “Joule heater” that produces 10-20°C “heat” 0.000’004% are used for information and 99.999’996% create cooling demand (negative heat)

Consequence: Reduce resistances and increase density!

PUE = 3.3

Brouillard, APC, 2006

Heat pump exergetic efficiency: ~25%

Compressors (Refrigeration)

Chilled Water CRAC/CRAHs

Condenser

Racks & Fans

Electrical Power

Chiller 33%
Humidifier 3%
CRAC 9%
IT Equipment 30%
PDU 5%
UPS 18%
Switchgear / Generator 1%
Lighting 1%

Invest 5%

Compression

Reuse

Cooling towers
Chip-cooling improves efficiency AND carbon footprint
- Cooling with $\Delta T = 20$ instead of 75°C, saves 50% energy
- Reuse: ~2000 Homes with 10 MW datacenter

Carbon footprint reduction in all climates
- Hot climates: Free cooling, desalination

Europe: 5000 district heating systems
- Distribution of 6% thermal demand
Datacenters with Solar Energy and Heat Reuse

Phase I (2012)

IDPX DWC dx360 M4
9288
HPPlpack Performance
2.9 PFLOPS
Power Dissipation
up to 3.6 MW

World’s Most Powerful & Energy Efficient x86 Supercomputers
Power Usage Effectiveness
PUE 1.1

200 PFlops
GPU+CPU
Summit (2018)
Oak Ridge NL

6-7 PFlops
Phase II (2015)
CPU only

IDPX DWC dx360 M4
9288
HPPlpack Performance
2.9 PFLOPS
Power Dissipation
up to 3.6 MW

World’s Most Powerful & Energy Efficient x86 Supercomputers
Power Usage Effectiveness
PUE 1.1

200 PFlops
GPU+CPU
Summit (2018)
Oak Ridge NL

40 Racks
CoolMUC-2 (2016)

Density increase through water cooling combined with power delivery
⇒ Density: Key for efficiency

1000x denser and 10x more efficient!!
Density enabled by reduced thermal & electrical Resistance

Heat-driven datacenter cooling

E_reuse

240 Racks
CoolMUC-2 (2016)

Phase II (2015)

NXS DWC nx360 M5
3096
HPPlpack Performance
2.8 PFLOPS
Power Dissipation
up to 1.3 MW

World’s Most Powerful & Energy Efficient x86 Supercomputers
Power Usage Effectiveness
PUE 1.1

200 PFlops
GPU+CPU
Summit (2018)
Oak Ridge NL

How to become Physically Net Zero by 2030, Bruno Michel, bmi@zurich.ibm.com © IBM Research 2021
Bell’s Law: More Integration

- Every 12-15 years restart new generation
- Hardware cost fraction decreases from 100% (mainframe) to <10% adding functionality
- Sensing and communication miniaturized
- Low thermal/electrical resistance enable density
- Sensing and computing meet in wearables
- **Remember: proximity improves efficiency!**
- Efficiency and low cost due to Bell’s law

- **10x denser and 2x more efficient!!**
- **1000x denser and 10x more efficient!!**

Technology developed with …

---

**Volumetric Density Scaling Impact**
- 5'000x smaller power
- 50’000’000x denser
- Scalability to zetascale

Cloud Zero Carbon Emission Design

Bad airflow and static cooling cause PUE > 1.5

Rear Door Heat Exchangers
and best practices: cold/hot aisle etc.
Reduces PUE=ERE = ~1.1
Rest of Cloud is here

Most Cloud is here

Greenwashing

Cloud CFP = 110 ... 120%

CO₂

CO₂

Hot Water Energy Reuse
Power Usage Effectiveness = 1.1
Energy Reuse Effectiveness = 0.7
Selling energy for reuse (increased RoI)
Summit

Cooling Reuse with Warm Water

Cloud could leap to here
CFP = 0% (without RECs)

Cloud leap to here
Carbon Footprint = 0%
Negative Carbon Footprint (without RECs)

Heat reuse + renewable energy
ERE = 0 → CFP = 0
Physical offset with zero carbon intensity equivalent to negative emission
Needs new metric to drive industry to this goal!
SuperMuc

Integrated power and cooling
reduces energy demand (E_{IT}) by 10 ... 10'000

CFP = E_{IT} × ERE × CI

2. Stepwise introduce chip-level hot water heat reuse after datacenter modification

….. Stairway to paradise .....
High performance microchannel coolers

Heat exchanger

Pump

Underfloor heating

Economic value of heat reduces datacenter total cost of ownership by 50-70% lower energy cost

60°C

65°C

>700 W/cm²

Sustainable generation
Electricity and heat

Sustainable generation

Fresh water scarcity

Renewable heating and cooling

Zero-emission datacenter

High-concentration PV/thermal

Adsorption heat pump
Carbon capture

Membrane distillation
desalination

Electrochemical redox energy conversion

Sustainability Impact: 5 Gt/yr (out of 37 Gt/yr)

How to become Physically Net Zero by 2030, Bruno Michel, bmi@zurich.ibm.com

© IBM Research 2021
Edge Computing Trends

$3.2B$: U.S. edge computing market size in 2025

27%: Annual global growth rate for edge computing through 2023

33%: Service provider network capacity within a metro network by 2022

45%: Percent of IoT data acted on at the edge in 2020

Examples of enterprise edge applications:

“The edge is where your customers are”

Machine control  Equipment monitoring
Patient monitoring  Robot surgery
Environmental monitoring  Vision-based analytics
Remote facility monitoring  Security services
Farm monitoring  Asset tracking
Fleet vehicle diagnostics  Autonomous delivery

5G and Edge Computing function separately, but deliver enhanced benefits when combined
Human Centric Sensing and Computing Strategy

Context key for relevant personalized cognitive services in wellbeing and work safety

IoT / wearables revolutionize healthcare for chronic diseases and elderly care by enabling **data-driven preventive medicine**

Today
- Single channel human reality
- Other channels inaccessible

Future
- Multichannel human reality
- All channels accessible

(Cognitive edge computing & companion)

Good internet context BUT
Severely limited local context
(limited relevance of cognitive services)

(Edge) Application driven growth

Context rich cognitive services
Human Centric Sensing and Computing

- Physiology / Medicine
- Psychology / Behavioral change
- Acceptance / Usability / User experience

modular platform adapts to use cases and strengthens base of AI pyramid

Humans as largest data source and consumer of AI

Artificial Intelligence

Big Data Automation

Data Collection/Preparation

Human in the Center

- Cloud Services
- Cloud Hub
- Sensors
- GUI
- User Interface
- Data Hub and Sensors
- Use Cases

API

Conversational
Inter-person Analyst.

How to become Physically Net Zero by 2030,
Bruno Michel,
bmi@zurich.ibm.com
© IBM Research 2021
Wellbeing Pilot Functionalities and Datasets

**Functionality of Wellbeing Pilot**
- Monitor elderly, workers, patients, or athletes
- Service provided by care-givers, captains/supervisors, or coaches
- Privacy and GDPR need limited data lifetime
- Elderly Care data from ActivAge
- Work safety Data from Firefighter trainings
- Data on monitoring and coaching of people with chronic diseases
Platforms and Power Measurement

Platforms and tools for power measurements:

- Movidius NCS2,
- Coral Dev board
- Coral Dev Sick,
- Jetson Nano,
- Raspberry Pi4,
- Joy-IT UM25C,
- SmartMe.

- Baseline to isolate inference or training energy and improve accuracy:
  - Isolated from inference of
    - MobileNetV2 (yellow)
    - ResNet-50 V1 (brown)

- Assuming background processes invariant

- Recognition gradient \( R_{\text{grad}} \) difference probability best to second-best class
- Probability of anomaly minus probability of normal
- Search results gradient calculation is similar
- Determined from 100-1000 test cases
- When true label is not found gradient is subtracted
Energy precision ratio $M$ for $\alpha=4$ emphasizes accuracy over power consumption.

\[ M = \text{Error}^\alpha \times EPI \]

- Recognition efficiency shown for the same models and platforms.

\[ RE = \frac{\text{REC}_{\text{grad}} \times CI}{\sqrt{E_{\text{inf}}}} \]

**Reduce** $E_{\text{move}}$ and $E_{\text{proc}}$
Lifecycle Efficiency

- Lifecycle recognition efficiency for different models (color) and different platforms (line type)
- Needs many uses of the model (F) to amortize training energy
  Winner mobile net on TPU
- Cloud suffers from data transport energy overhead
- Transition training to inference domination beyond 1 Million uses

\[
R_{ELC} = \frac{REC_{grad} \times Cl}{\sqrt{E_{inf} + E_{train}/F}}
\]
Recognition Efficiency

\[ \text{RE} = \frac{\text{REC}_{\text{grad}} \times \text{Cl}}{\sqrt[2]{\text{Einf}}} \]

- Simple to understand ratings A-D
  - A rating: \(100 < \text{RE}\)
  - B rating: \(10 < \text{RE} < 100\)
  - C rating: \(1 < \text{RE} < 10\)
  - D rating: \(\text{RE} < 1\)
- Simple calculation method
- Balanced influence from accuracy, energy, and number of classes
- Always based on Top 1
- Used with \(\text{REC}_{\text{grad}}\) but also accuracy
Green Datacenter: Green AI

AI energy problem
Energy for deep learning doubled every few months, resulting in 300,000x increase from 2012 to 2018

Step 2: Compare IR AI models on same system
Recognition Efficiency varies >50x for models on same hardware
- Optimize network model selection
- Once for all training; lean libraries

Step 3: Rate different complexity models (biometric, image, sound) (such as STAR rating)
→ Standardized measure of Green AI
→ Bar rises as new accelerators develop
→ Assess whether a model is mature
→ Assess label quality in dataset

Step 4: Compare model on different systems
→ Hardware Acceleration and lean libraries
- Low precision ASICs 8-2 bit
→ Tunable precision and dynamic delegation
→ Analog accelerators

Step 1: Universal Metric: Recognition Efficiency
Improvement need measurement!
- \( \text{RE} = \text{Accuracy} \times \text{Cl} / (E_{\text{inf}})^{0.5} \)
- Power measurements on edge
- Compare models on different systems
- Used as standard metric to define what is green or red AI
Cooling Supercomputers using Adsorption Heat Pumps

**Announced February 2018**

Adsorption chilling applied to next-gen LRZ, generating 600 kilowatts of chilled water.

**Return-on-investment**

- **Cost of retrofit** ~100 kEUR
- **Piping, HEX, chillers**
- **Energy savings** ~10 kEUR/year
  - Compared to operation at 40°C cluster inlet temp.
- **Payback time:** 10 years

**LRZ and Fahrenheit AG build data center specific adsorption chillers with improved efficiency at lower temperature levels.**

**Profitability of zero-emission datacenters in HPC and in Cloud requires thermally driven heat pumps with 3 - 10 x lower CAPEX and 2 - 3 x better exergetic efficiency**

⇒ **larger value of heat**
Technology shown 1878 first time (silica gel and water)
Sorption cooling mature. Ammonia sorption heat pumps were replaced by electrical heat pumps with ozone killing FCKW working fluids
Our addition improves technology 50’000 times

Medical oxygen concentrators (COVID Pandemic)
Rapid pressure swing is a huge success story!
Key overlap with Rapid Pressure Swing Technology: Structured materials with improved mass transport
(Slow) Thermal Swing process is established in gas purification but not gas separation
Heat Driven Heat Pumps or Adsorption Chillers

**Adsorbents**
- Silica gel
- Zeolite
- Activated carbon
- Alumino-phosphates
- MOFs
- Salt composites

**Refrigerants**
- Water, Methanol, Ammonia, Sulfur Dioxide, Carbon Dioxide

**Transport dynamics**
- Improve adsorption capacity
- Improve dynamic utilization
- Select for application

Thermal resistance reduction improves efficiency of sorption chillers from 20% to 50%!!!
Rapid Thermal Swing Technology in Heat Pumps

Adsorber unit-cell

Research System
TRL 4-5

First 1 kW
System TRL 5

Second 1 kW

Third 10 kW
System TRL 6

HEX TRL 6

Unit Cell

1 kW System TRL 4-5

3x

10 kW AdHP demonstrator

1 kW System TRL 7
Capture Carbon in Power Stations

**Post Combustion**
Retrofit to power plants. Pluck CO₂ from flue gas: amine scrubbing, ammonia and limestone membranes selectively bind and release CO₂, followed by dehydration, compression, and storage.

**Pre-Combustion**
Air separator feeds O₂ into gasifier, baking coal at 700°C to release H₂, CO, and CO₂; syngas. Shift reactor converts to H₂ and CO₂. Separation of H₂ and CO₂.

**Oxyfuel combustion**
Burns coal/gas using pure O₂, from an air separation unit. Purification unit removes pollutants and compresses CO₂.

Images: Vattenfall Factsheet 1332
Rapid Thermal Swing Adsorption Gas Separation

Gas separation expensive (Amine scrubbing is $30-60/t) due to OPEX
- Energy for regeneration (30% of PS)
Pressure swing adsorption (PSA) $40-63/t not viable due to CAPEX and OPEX BUT RPSA 3 min cycles >10x better productivity
- Lower CAPEX but same OPEX (>25% of PS)
Temperature swing adsorption (TSA) 2x worse than PSA due to 1h regeneration ➔ OPEX

Rapid Temperature Swing Adsorption (RTSA) based on adsorption chiller
- IBM’s RTSA uses low temperature waste heat driven re-generation
- Reduced cycle time with structured sorbents for best mass/heat transport
TSA to RTSA uses similar principles than PSA to RPSA
- Reduced CAPEX and OPEX
- ~3 min cycles reduce CAPEX 20x over TSA and 5x over PSA
RTSA reduces CAPEX/OPEX >10x and improves lifetime vs. PSA, TSA, and membrane processes
RTSA reduces gas separation cost to <<20$/t after scale-up, makes it economical with short ROI in EU
Summary: Physically Net Zero GHG Emissions by 2030

- **Split Sustainability Problem to Solve ... Efficiency is Key!**
  - Physical offset of waste heat, PUE insufficient metric ...
  - Cloud vendors refused innovation from HPC
  - Dense computers ~99% energy needed for data movement, ~1% for compute

- **Renewables from Same Grid Section ... NO RECs**
  - Thermal resistance to make renewables more efficient
  - Supply following loads

- **From Cloud to Edge ... Proximity Improves efficiency**
  - After Cloud benefits have been exploited there is no way around edge ... “pendulum”
  - Privacy, resilience, efficiency ...

- **Stop Red AI (Energy waste) ... Start Green AI**
  - Universal Metrics ... You only improve what you measure! (Recognition Efficiency like Energy Star)

- **Carbon capture removes residual emissions ... NO financial offsets**
  - Financial offsets are 100x inferior (cheaper) than physical offsets ... and thus an “innovation killer”
  - Instead paying financial offsets invest into carbon capture breakthrough innovation!!
Key Messages

- Financial Offsets and RECs are “Innovation Killers”
- Invest into carbon capture breakthrough innovation
- “Stairway to heaven” datacenters with physical offset
- To consume renewables, supply following loads needed
- Physical Offsets 100x more valuable and important
- Start Green AI with Efficiency Measurements
- Need for Universal Metrics (like Energy Star)
Thank you for your attention!
Publications


R.P. Luijten, A. Doerig, S. Paredes, T. Engbersen, F. Buining, B. Michel, ”Objective, innovation and impact of the energy efficient DOME MicroDataCenter”, 2017 Int’l Conf. on Advances in Computing, Communications and Informatics (ICACCI), 10.1109/ICACCI.2017.8125958


Publications (2)


