Data Centers Get Serious:

Unlimited Demand Meets Practical Reality

(http://www.esp.org/briite/meetings)

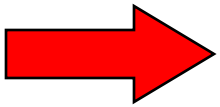
Robert J. Robbins

rrobbins@fhcrc.org

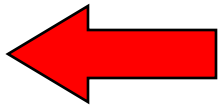
(206) 667 4778

Data Centers Get Serious:

Unlimited Demand Meets Practical Reality



(http://www.esp.org/briite/meetings)



Robert J. Robbins

rrobbins@fhcrc.org

(206) 667 4778

Abstract

By some estimates, more than half of all organizations will be facing a shortfall in data-center capacity in the next twelve months. Biomedical research organizations are no exception. Sequencers are now available than can produce 1.5 terabytes of data per run. Companies like Pacific Biosciences are developing single-molecule, real-time approaches to sequencing that hold out the promise of producing a billion bases of sequence every five minutes. That translates into a 15x coverage of the human genome in less than four hours — the entire sequencing effort of the human genome project starting at breakfast and done before lunch. The storage, management, and analysis of such data flows will require prodigious computing power and staggering mass-storage capacity. Individual RO1 grants will easily require tens, or hundreds, of terabytes of disk space. It is now routine for institutions to offer dedicated 250-node compute clusters to recruit individual computationally intensive faculty.

Information technology is critically important for the success of biomedical research institutions, but the demands on our data centers are rapidly outstripping our ability to deliver. What to do? Determining how best to meet the wave of data-center demand will be the topic of this meeting — Data Centers Get Serious.

Topics

- Driving Factors
  - Moore's Law / Ubiquitous Computing
  - Post-Genomic Science / New Technologies
- Possible Solutions.
  - Build More
  - Use Less (virtualization / green computing)
  - Use Someone Else's (COLO, Outsource, Cloud)

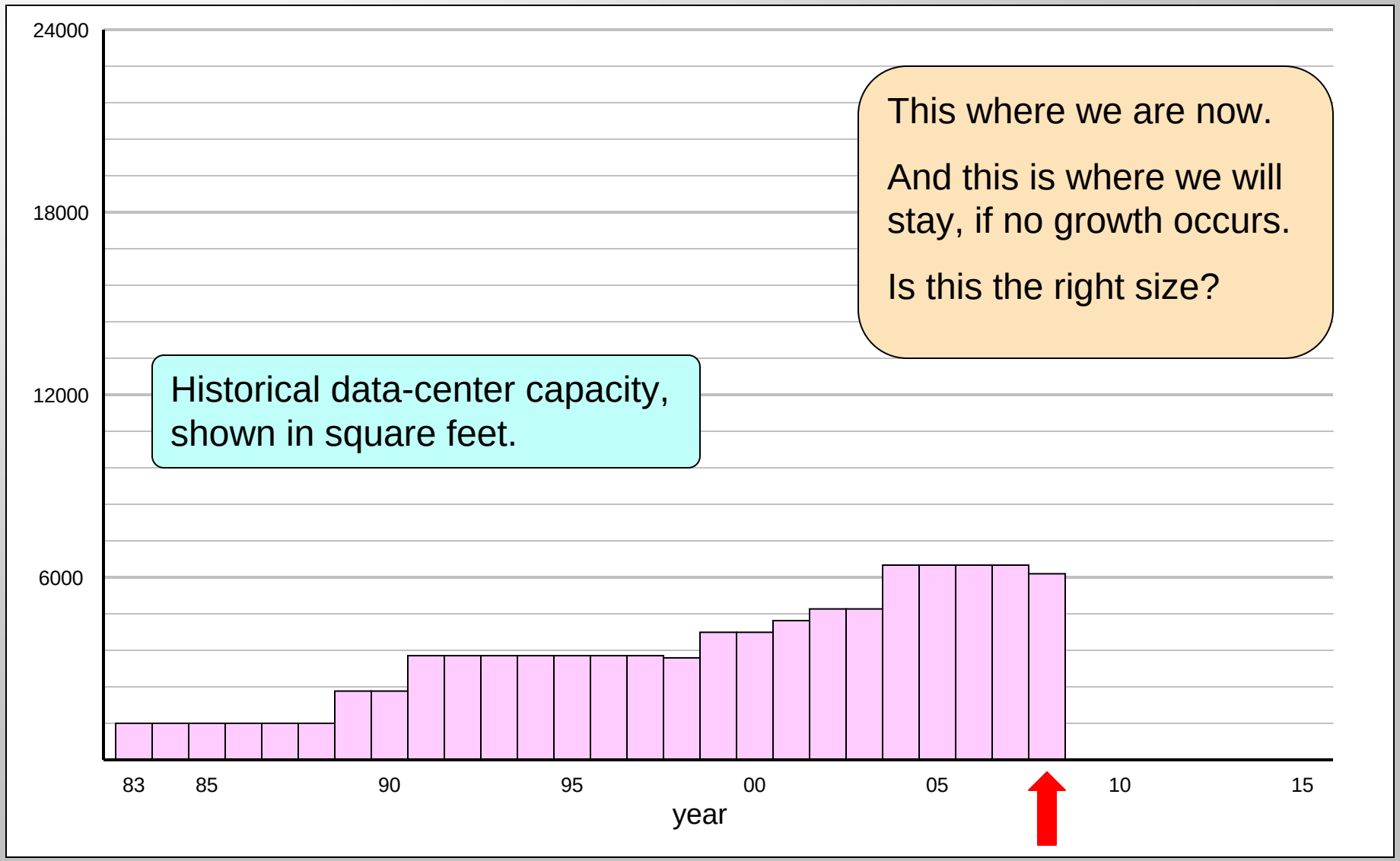
The Problem

- Demand for computing is increasing exponentially, while data center capacity is relatively static.
- We are outgrowing our ability to house and manage the hardware infrastructure for our information systems.
- This is a problem for everyone.
- But it is especially a problem for biomedical research:

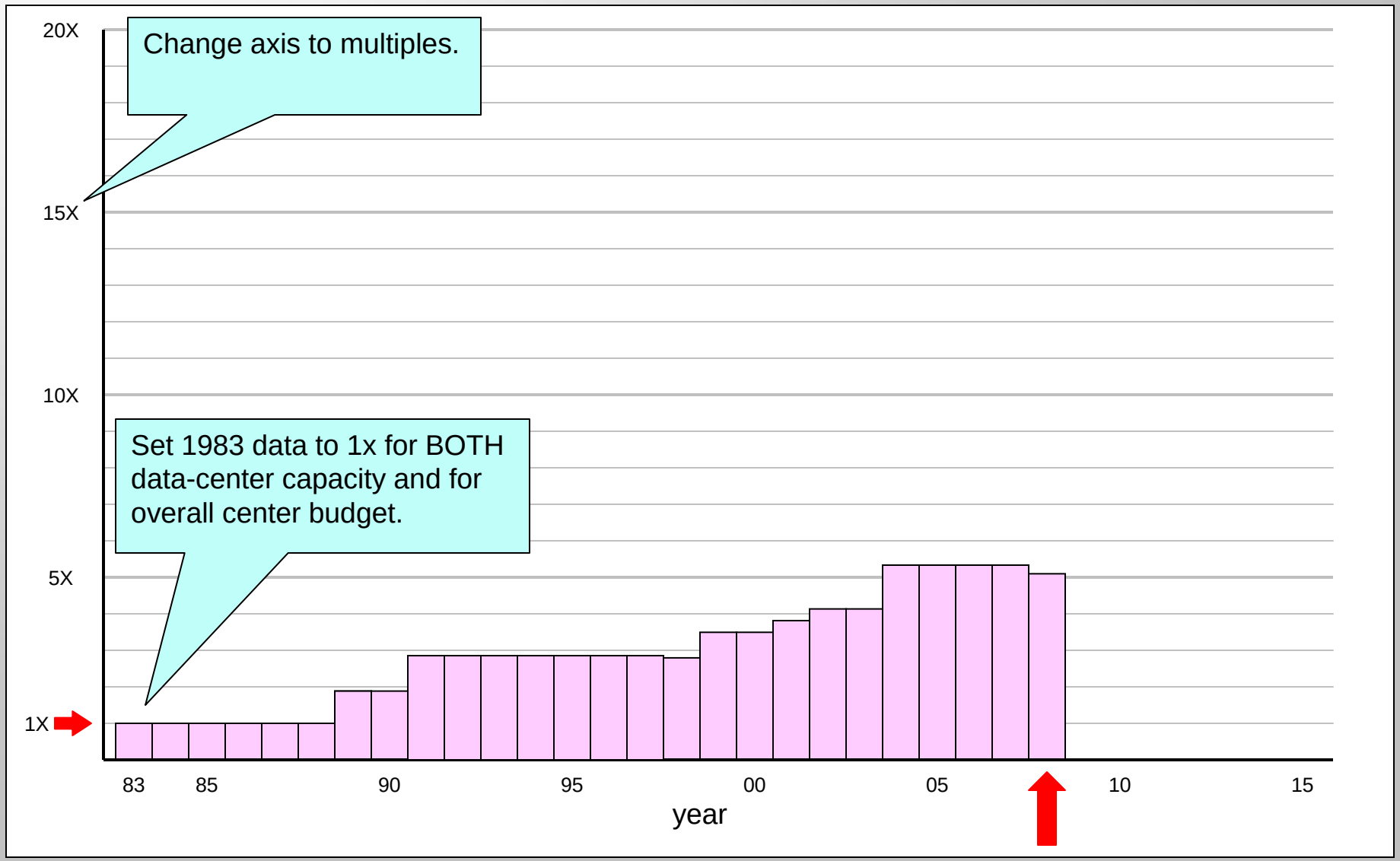
  Moore's Law is driving advances in computing and advances in computing are driving advances in biomedical technology.

FHCRC Experience

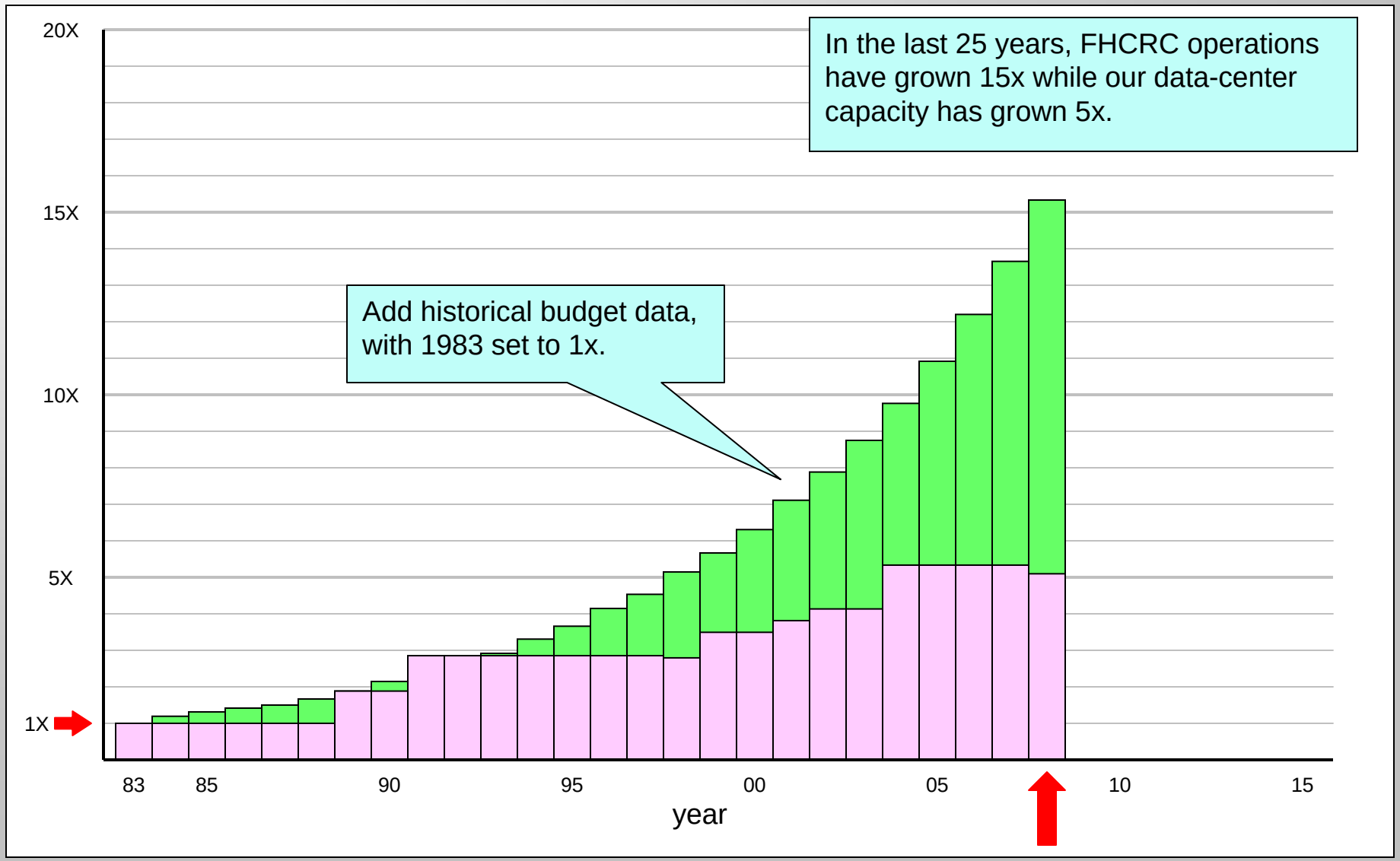
Growth at FHCRC



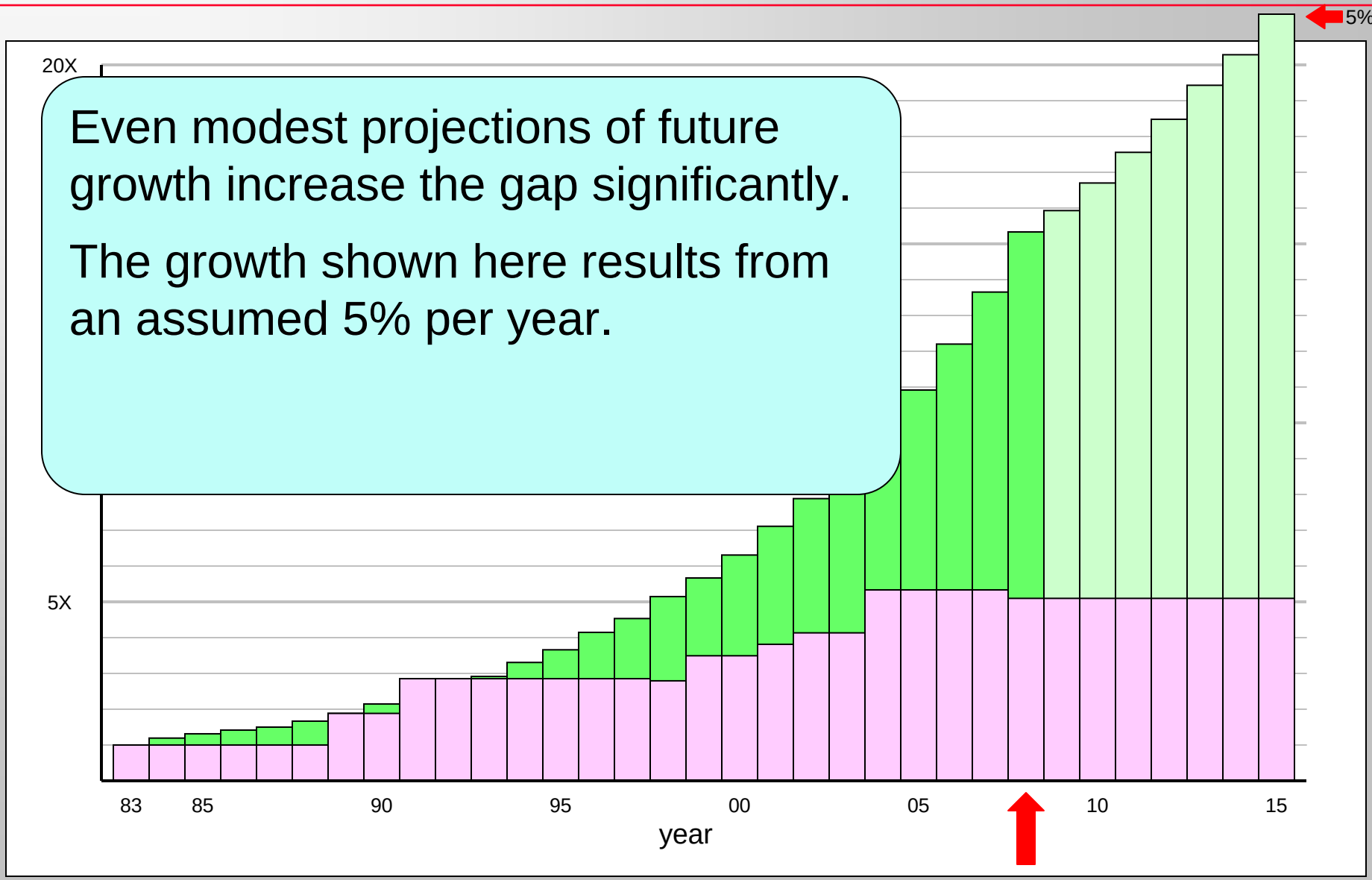
Growth at FHCRC



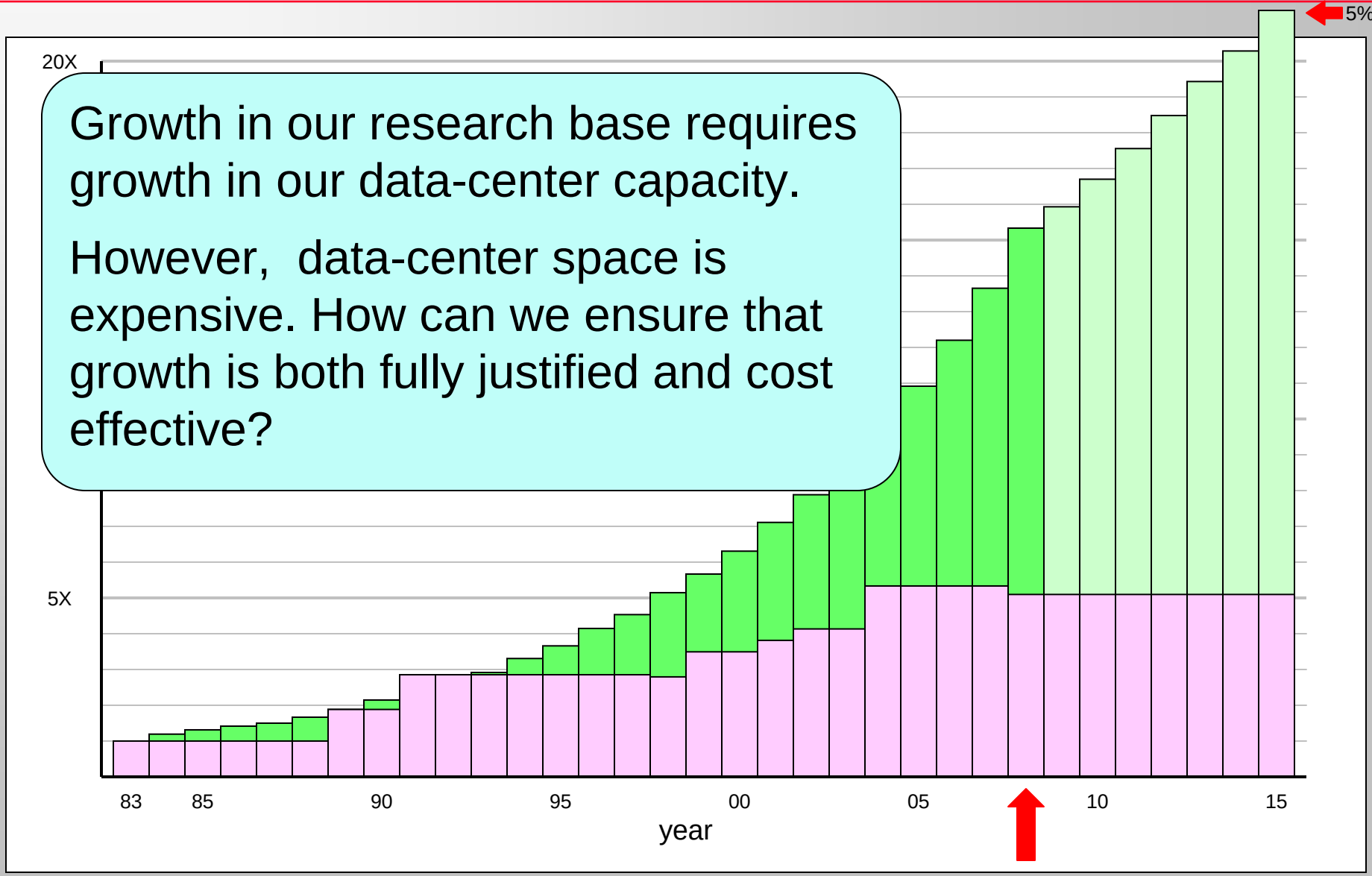
Growth at FHCRC



Growth at FHCRC



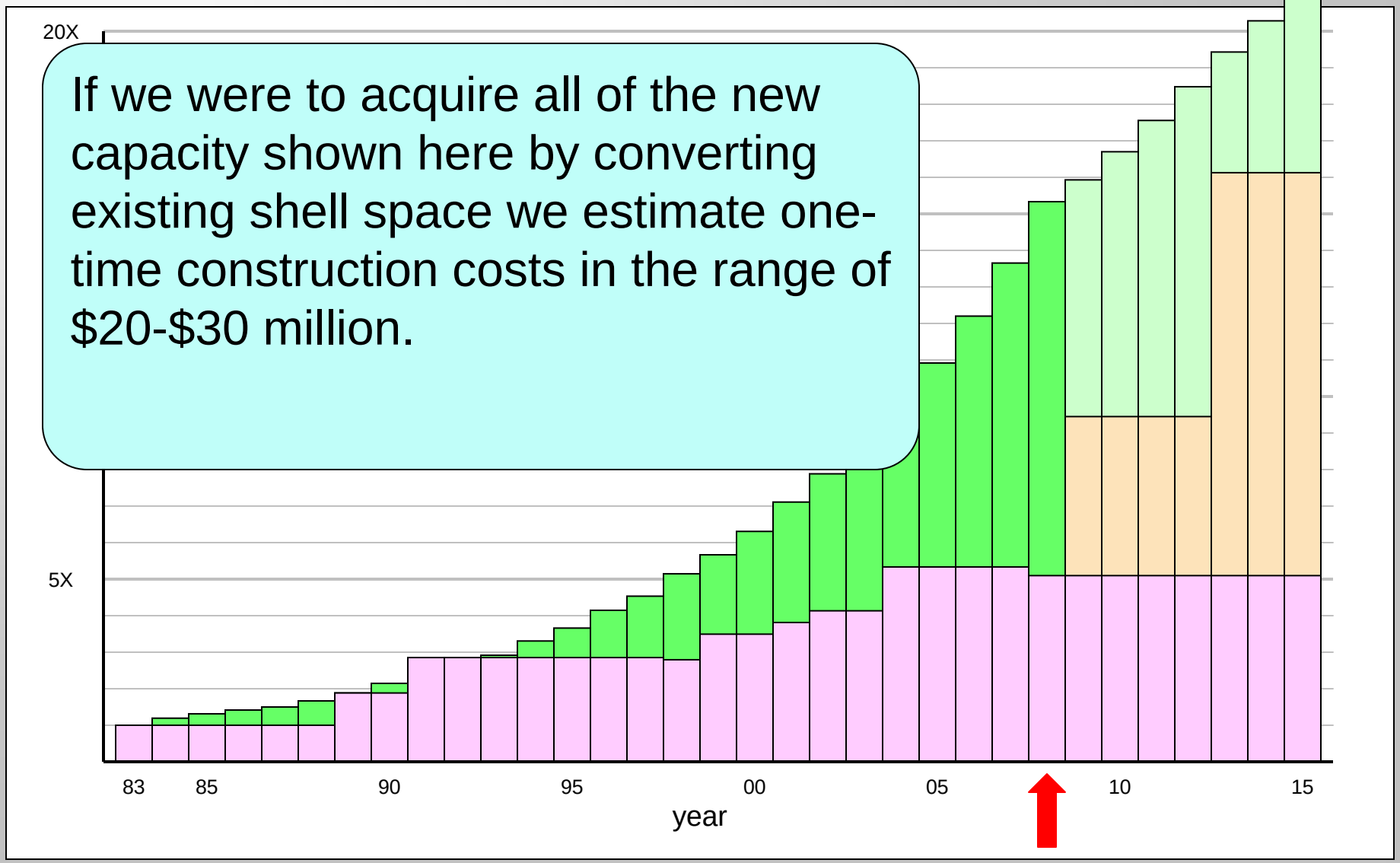
Growth at FHCRC



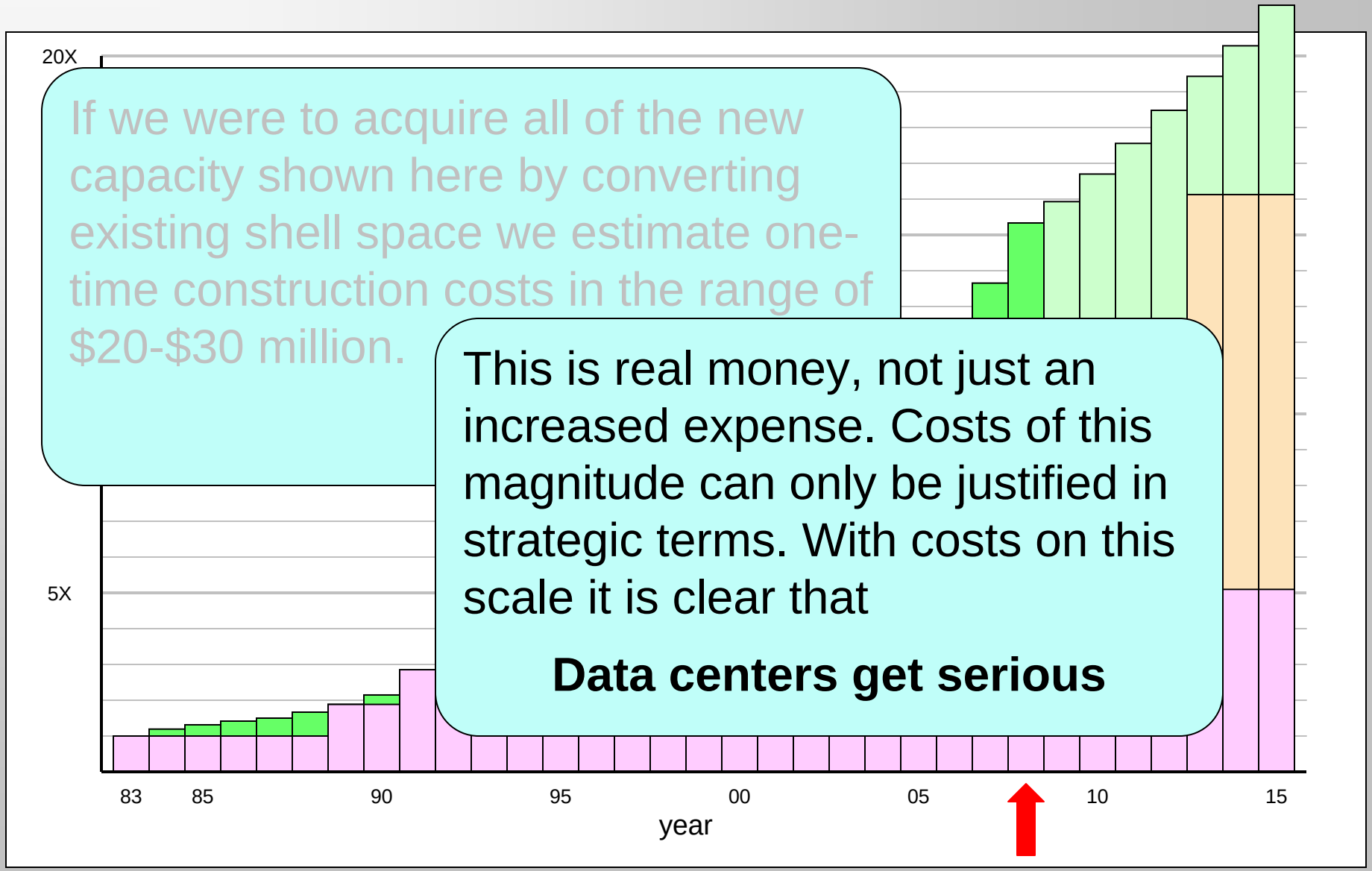
Growth in our research base requires growth in our data-center capacity.

However, data-center space is expensive. How can we ensure that growth is both fully justified and cost effective?

Growth at FHCRC



Growth at FHCRC



Driving Factors

Moore’s Law

Moore's Law

The Law:

- The number of transistors that can be placed on a chip doubles approximately every 18 months.

Gordon Moore, 1965

Moore's Law

The Results:

- The power of advanced computing continues to grow almost 60% per year.
- The price of constant computing continues to drop by almost 40% per year.

Moore's Law

Declining prices create new opportunities for computer use.

Cost at constant performance decreases exponentially.

Moore's Law

Declining prices create new opportunities for computer use.

These new opportunities push increasing demand beyond the ability of Moore's law to meet that demand.

Cost at constant performance decreases exponentially.

Moore's Law

Declining prices create new opportunities for computer use.

These new opportunities push increasing demand beyond the ability of Moore's law to meet that demand.

The result is that growing demand for computational services is overwhelming our ability to meet that demand.

Cost at constant performance decreases exponentially.

Moore's Law

Examples:

- In 1987, two gigabytes of mass storage (a trivial USB drive today) would have cost more than $30,000.

Moore's Law

Examples:

- In 1987, two gigabytes of mass storage (a trivial USB drive today) would have cost more than $30,000.
- Today, a single Powerpoint file is often larger than all of the hard disk capacity available on a typical personal computer in 1987.

Moore's Law

Examples:

- In 1987, two gigabytes of mass storage (a trivial USB drive today) would have cost more than $30,000.
- Today, a single Powerpoint file is often larger than all of the hard disk capacity available on a typical personal computer in 1987.
- Storing a single music album in good quality MP3 files requires more space than available on a typical 1987 computer.

Moore's Law

In 1987, a typical desktop computer was equipped with less than 40 mB of disk space.

Today it is hard to buy a new desktop computer with less than 80 gB of disk and some can easily have a 1000 gB or more. This is a huge increase in just 20 years.

Even greater growth challenges are occurring in our data centers.

Cost at constant performance decreases exponentially.

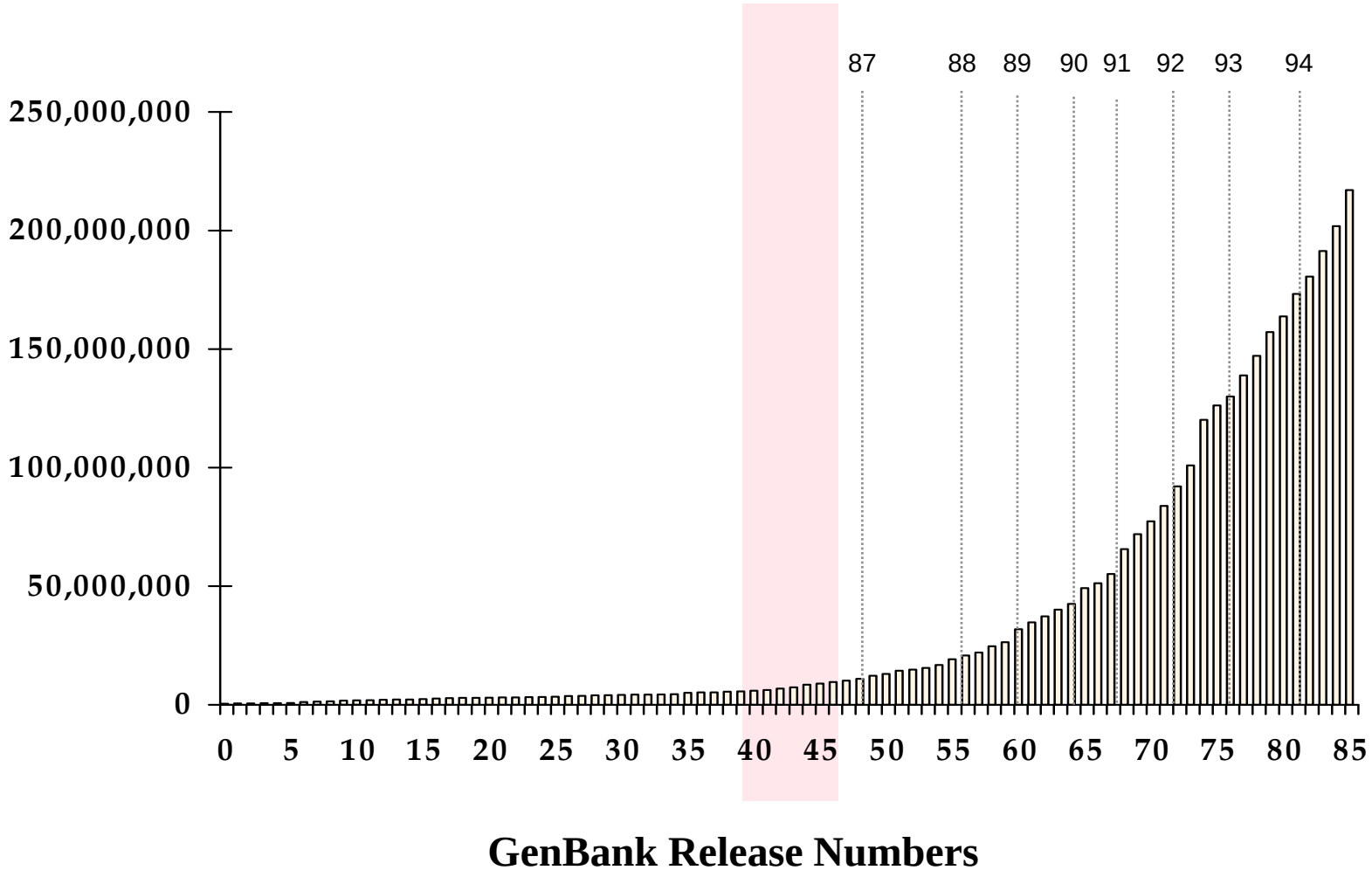
Driving Factors

Post-genomic Science & New Technologies

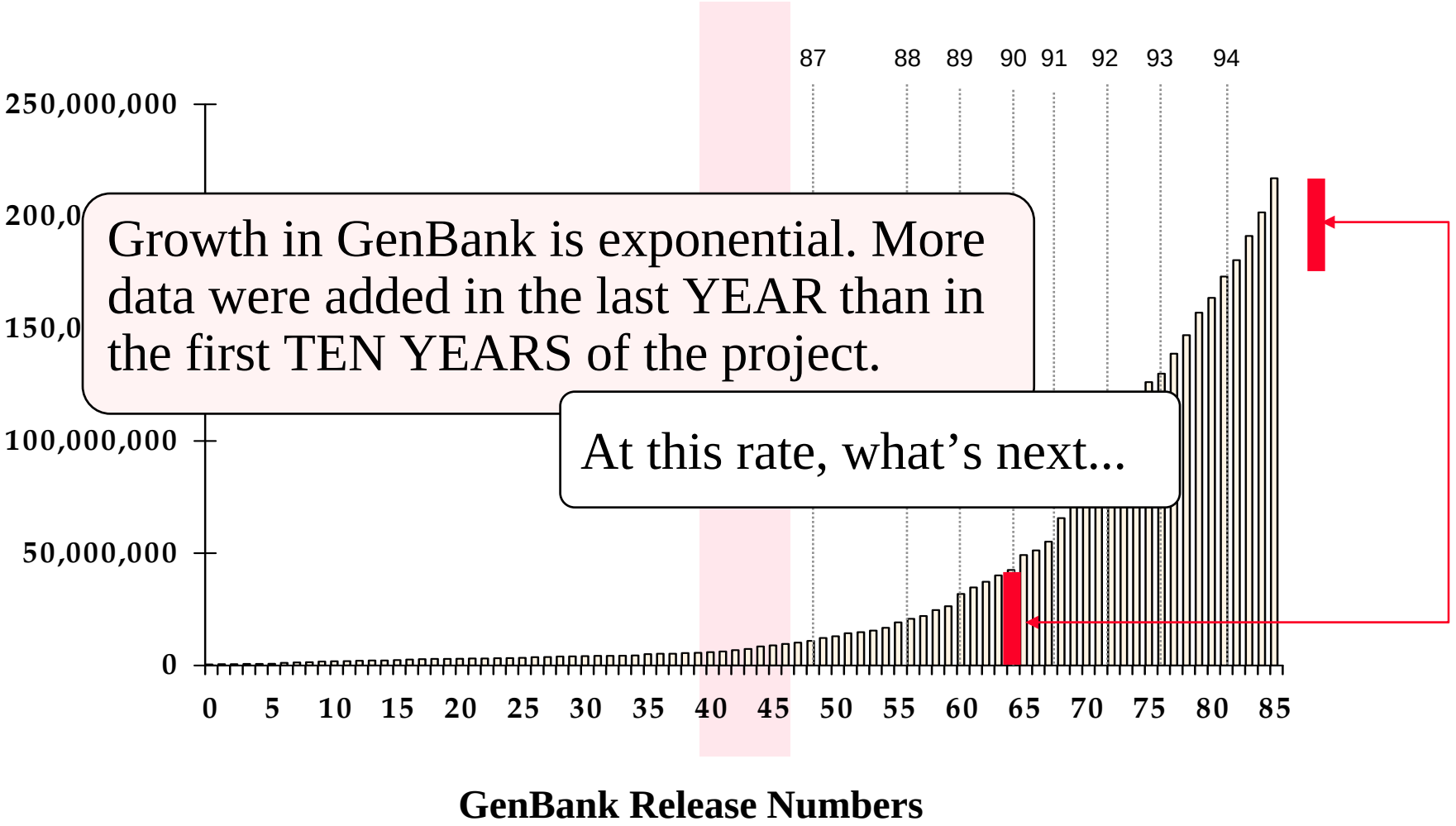
Awash in Data

Public Data Explosion

Base Pairs in GenBank (1994)

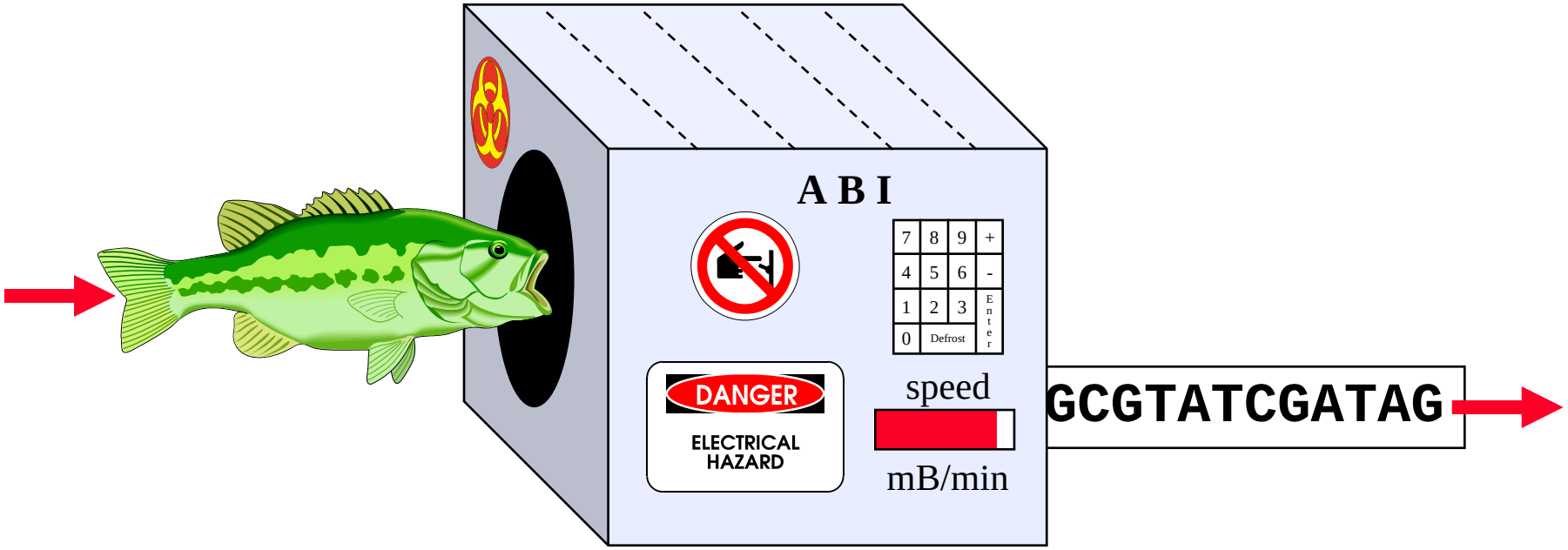


Base Pairs in GenBank (1994)



Samples to Sequence

A prediction from 1994…



In with the sample, out with the sequence...

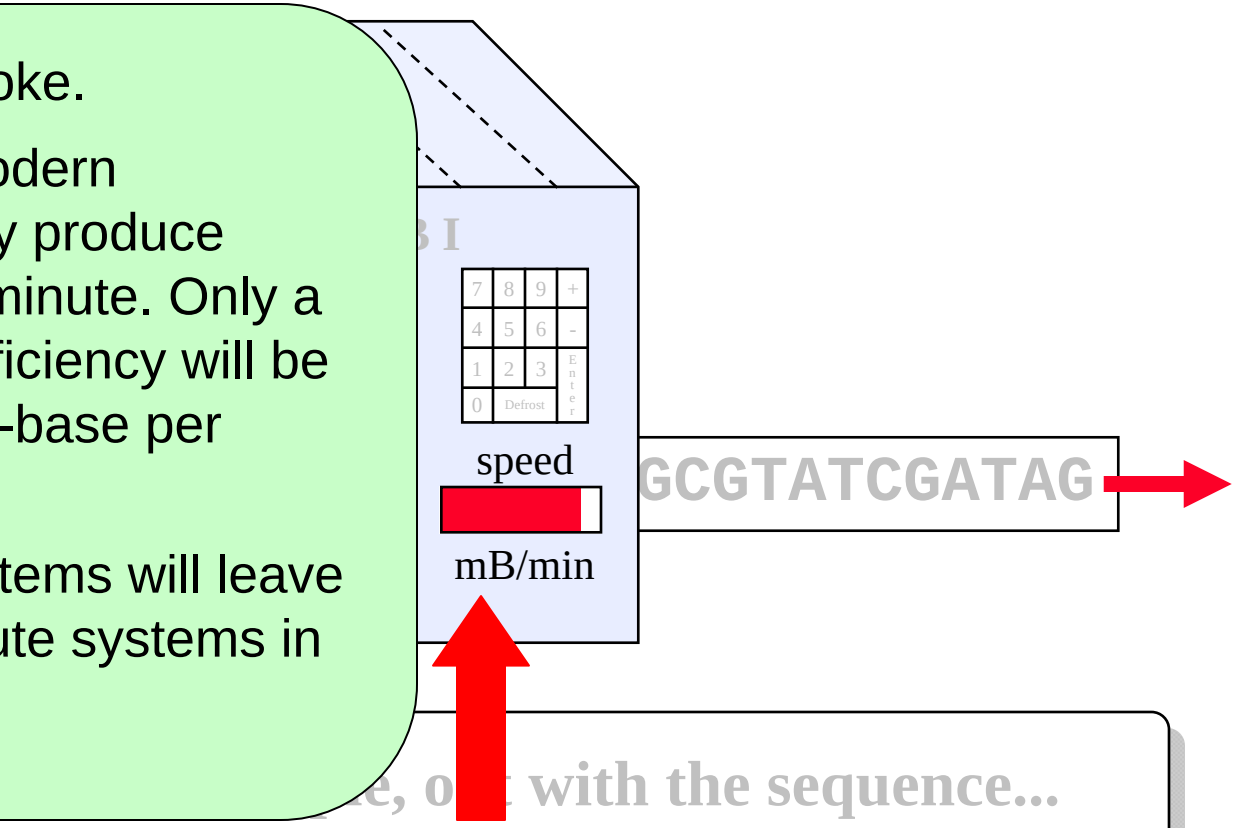
Samples to Sequence

A prediction from 1994…

This is no longer a joke.

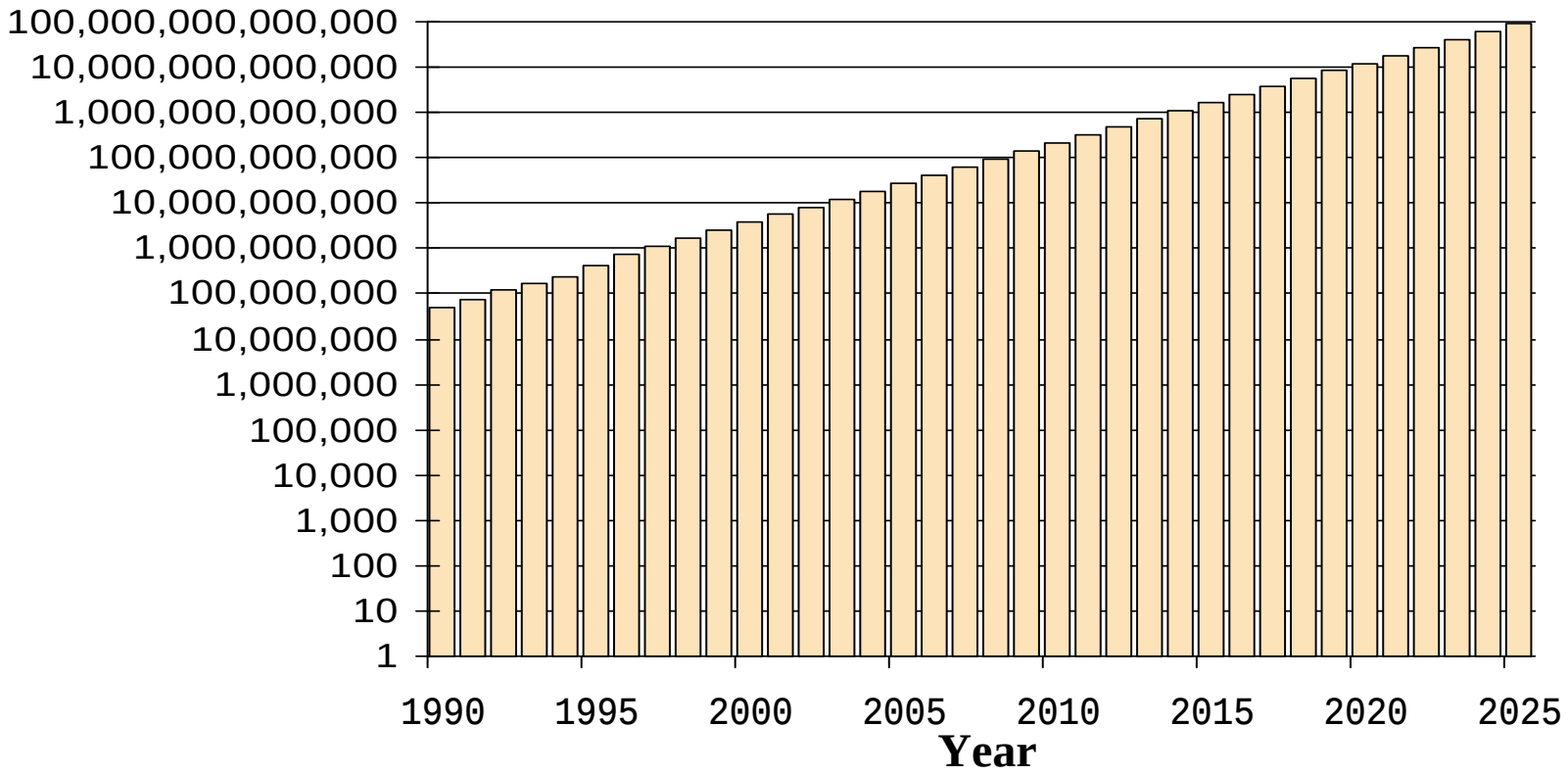
A state-of-the-art modern sequencer can today produce 500,000 bases per minute. Only a minor increase in efficiency will be required to hit mega-base per minute speeds.

Next-generation systems will leave mega-base per minute systems in the dust.



Projected Base Pairs in GenBank

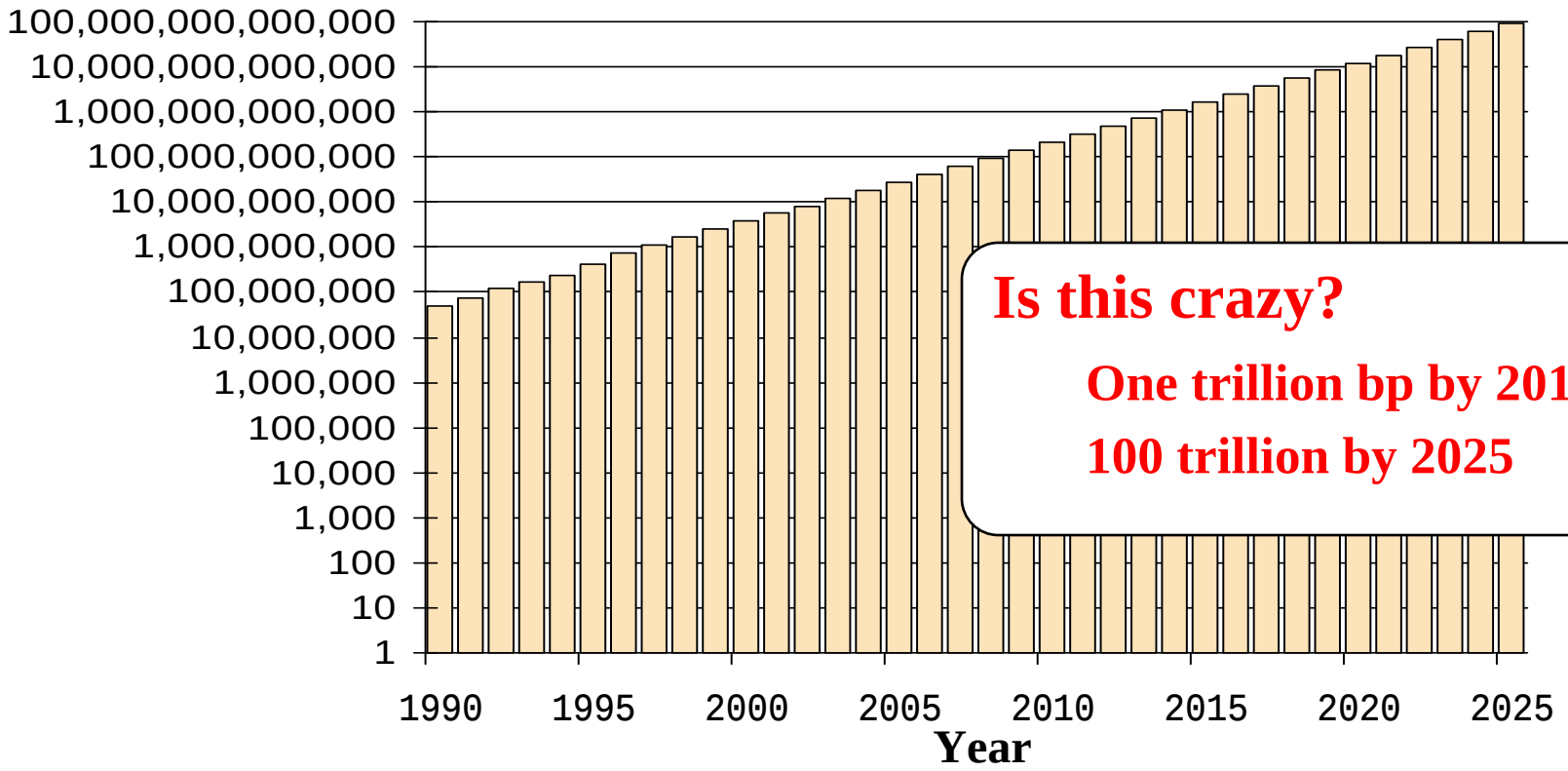
Another prediction from 1994…



Assumed annual growth rate: 50%

Projected Base Pairs in GenBank

Another prediction from 1994…



Is this crazy?

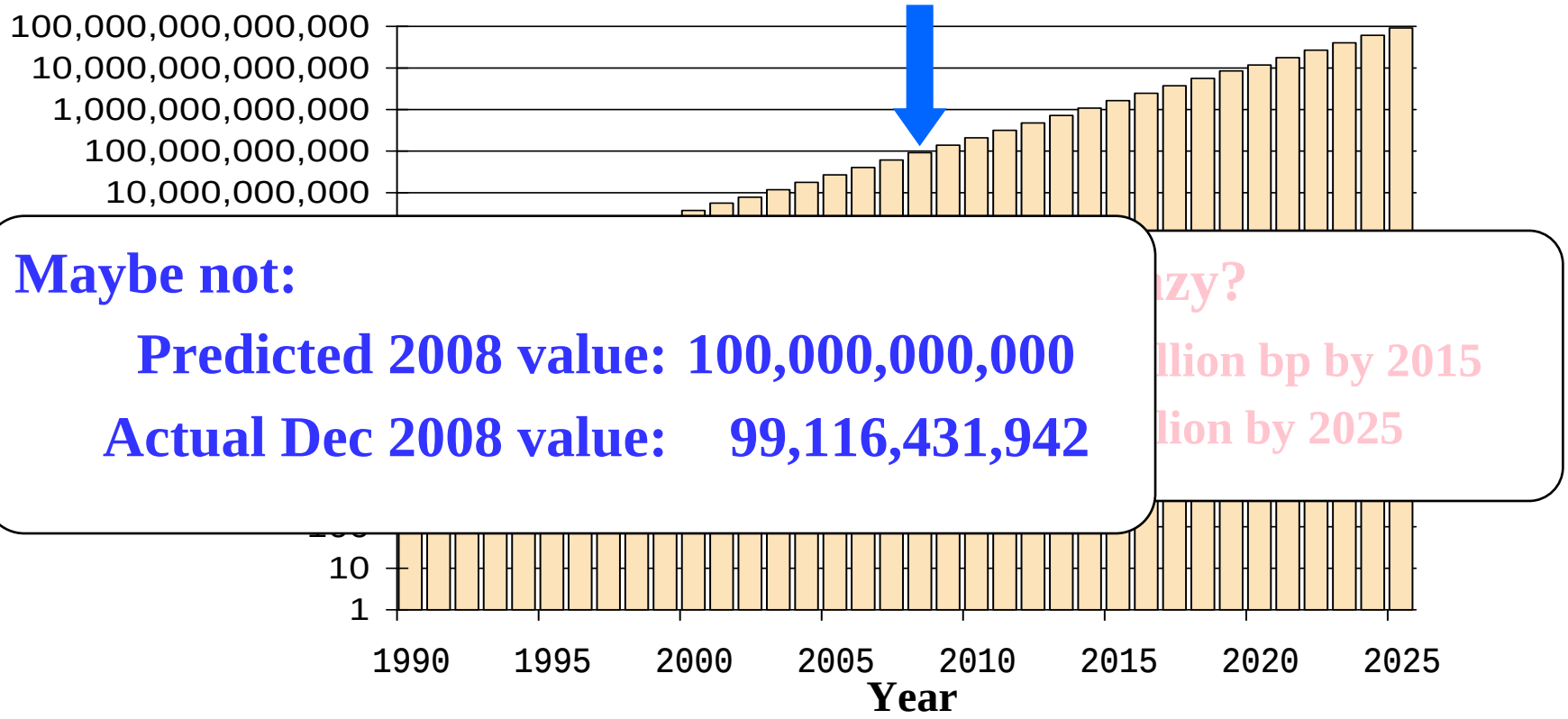
One trillion bp by 2015

100 trillion by 2025

Assumed annual growth rate: 50%

Projected Base Pairs in GenBank

Another prediction from 1994…



Maybe not:

Predicted 2008 value: 100,000,000,000

Actual Dec 2008 value: 99,116,431,942

…zy?

…llion bp by 2015

…lion by 2025

Assumed annual growth rate: 50%

Projected Base Pairs in GenBank

The point is,
sometimes even crazy
predictions do come true.

Projected Base Pairs in GenBank

The point is,
sometimes even crazy
predictions do come true.

Is there a limit?

Sequencing Speed

How fast can we read sequences?

ABI 377:	7,200	bases per hour
ABI 3700:	20,000	bases per hour
Illumina Solexa:	20,000,000	bases per hour
454 GS FLX:	20,000,000	bases per hour

Sequencing Speed

How fast can we read sequences?

ABI 377:	7,200	bases per hour
ABI 3700:	20,000	bases per hour
Illumina Solexa:	20,000,000	bases per hour
454 GS FLX:	20,000,000	bases per hour
Pacific BioSciences:	20,000,000,000	bases per hour

Sequencing Speed

How fast can nature read sequences?

One E. coli cell: 4,600,000 bases per hour

Sequencing Speed

How fast can nature read sequences?

One E. coli cell: 4,600,000 bases per hour

The rate at which one bacterial cell can replicate DNA is a good measure of how fast DNA can be manipulated at the molecular level.

All sequencing methods to date involve manipulating DNA at the molecular level, so this rate around 1,000 bases per second is probably a good estimate of the single-molecule speed limit for DNA processing.

Sequencing Speed

How fast can nature read sequences?

But that's a SINGLE-MOLECULE speed limit.

What happens if we work on more than one molecule at a time?

bases per second is probably a good estimate of the single-molecule speed limit for DNA processing.

Sequencing Speed

How fast can nature read sequences?

One E. coli cell:	4,600,000	bases per hour
10 E. coli cells:	46,000,000	bases per hour
100 E. coli cells:	460,000,000	bases per hour
1,000 E. coli cells:	4,600,000,000	bases per hour
10,000 E. coli cells:	46,000,000,000	bases per hour
100,000 E. coli cells:	460,000,000,000	bases per hour
1,000,000 E. coli cells:	4,600,000,000,000	bases per hour

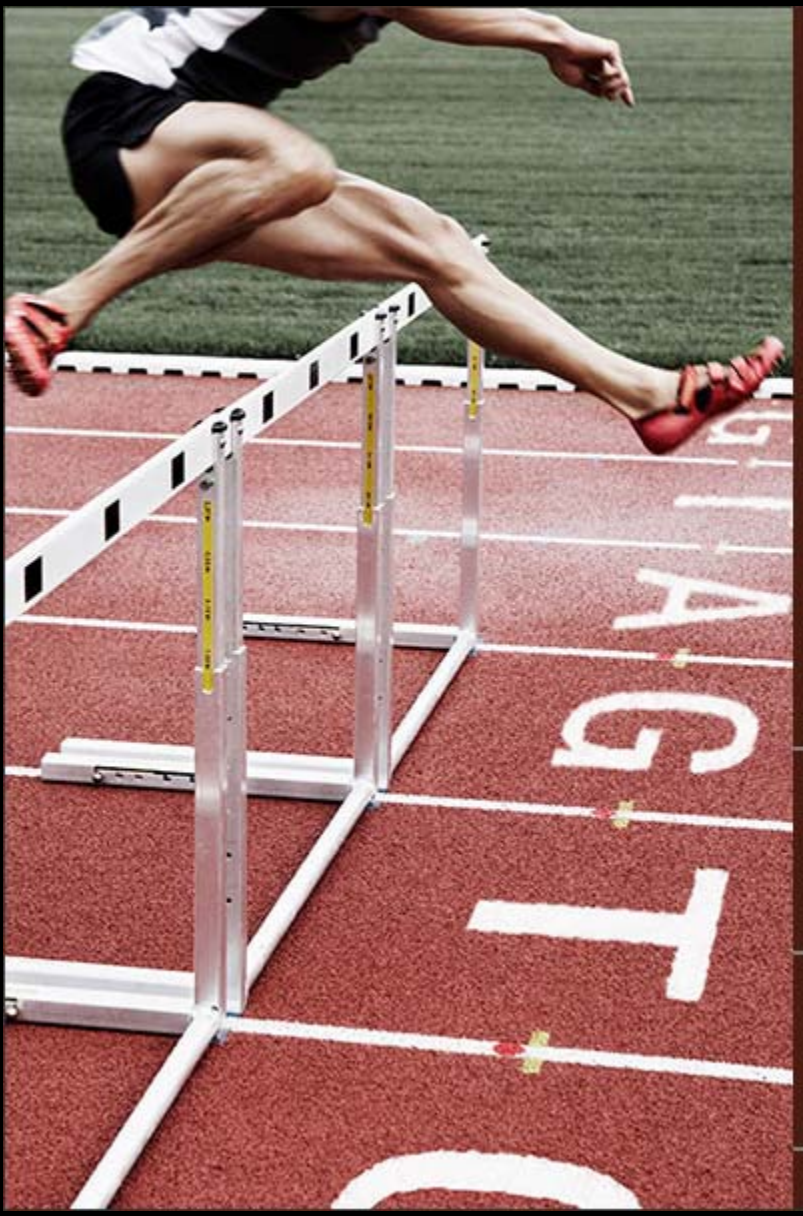
Sequencing Speed

How fast can nature read sequences?

One E. coli cell:	4,600,000	bases per hour
10 E. coli cells:	46,000,000	bases per hour
100 E. coli cells:	460,000,000	bases per hour
1,000 E. coli cells:	4,600,000,000	bases per hour

Working on 1,000 molecules in parallel would give about one copy of the human genome per hour.

Molecules are small, why not do 1,000 at once?



PACIFIC BIOSCIENCES™

Search

Register for More Information

ABOUT US | TECHNOLOGY | APPLICATIONS | NEWS & EVENTS | CAREERS | CONTACT US

Ready to leap…

A groundbreaking DNA sequencing technology is going to redefine the field. Single molecule real time.

Pacific Biosciences is a bold company developing a transformative DNA sequencing platform. Our breakthrough single molecule, real time (SMRT™) technology delivers the ultimate combination of long reads, low costs, and fast cycle times. A new paradigm for whole genome analysis is about to emerge.

NEWS

>> Intel, Others Back New DNA Sequencer

>> Genomes 'R' Us

EVENTS

>> The Genomics of Common Diseases 2008

SMRT™ TECHNOLOGY

View Our 4:05 minute Technology Demo

SITE MAP | PRIVACY POLICY | TERMS OF USE

PACIFIC BIOSCIENCES™

Search

Register for More Information

ABOUT US | TECHNOLOGY | APPLICATIONS | NEWS & EVENTS | CAREERS | CONTACT US

Ready to leap…

A groundbreaking DNA sequencing technology is going to redefine the field. Single molecule real time.

Pacific Biosciences is a bold company developing a transformative DNA sequencing platform. Our breakthrough single molecule, real time (SMRT™) technology delivers the ultimate combination of long reads, low costs, and fast cycle times. A new paradigm for whole genome analysis is about to emerge.

NEWS

- Intel, Others Back New DNA Sequencer
- Genomes 'R' Us

EVENTS

- The Genomics of Common Diseases 2008

SMRT™ TECHNOLOGY

View Our 4:05 minute Technology Demo

New Technologies

PacBio has solved this problem with the SMRT™ chip, which contains thousands of zero-mode waveguides (ZMWs). The ZMW provides the world's smallest detection volume, representing a 1000-fold improvement over existing single-molecule detection technology. Because the detection volume is so dramatically reduced, a single incorporation event can be observed against the background created by the high concentration of fluorescently labeled nucleotides. It makes possible the real-time observation of a single molecule of DNA polymerase as it synthesizes DNA.

New Technologies

PacBio has solved this problem with the SMRT™ chip, which contains thousands of zero-mode waveguides (ZMWs). The ZMW provides the world's smallest detection volume, representing a 1000-fold improvement over existing single-molecule detection technology. Because the detection volume is so dramatically reduced, a single incorporation event can be observed against the background created by the high concentration of fluorescently labeled nucleotides. It makes possible the real-time observation of a single molecule of DNA polymerase as it synthesizes DNA.

New Technologies

PacBio has solved this problem with the SMRT™ chip, which contains thousands of zero-mode waveguides (ZMWs). The ZMW provides the world's smallest detection volu ting sin det inc bac fluc rea pol

And each one of these zero-mode waveguides will carry out a sequencing operation on a single DNA molecule.

Hmmm. We are approaching the capabilities of thousands of individual E. coli cells.

Let's look at this in a little more detail…

New Technologies

Scale: 43.5 μm wide x 32.8 μm tall.

Each tiny PacBio SMRT chip contains thousands of zero mode waveguides.

New Technologies

TINY barely captures the size of the PacBio SMRT chip.

New Technologies

TINY barely captures the size of the PacBio SMRT chip.

Think of the letter O of IN GOD WE TRUST on a US quarter as a tiny bowl.

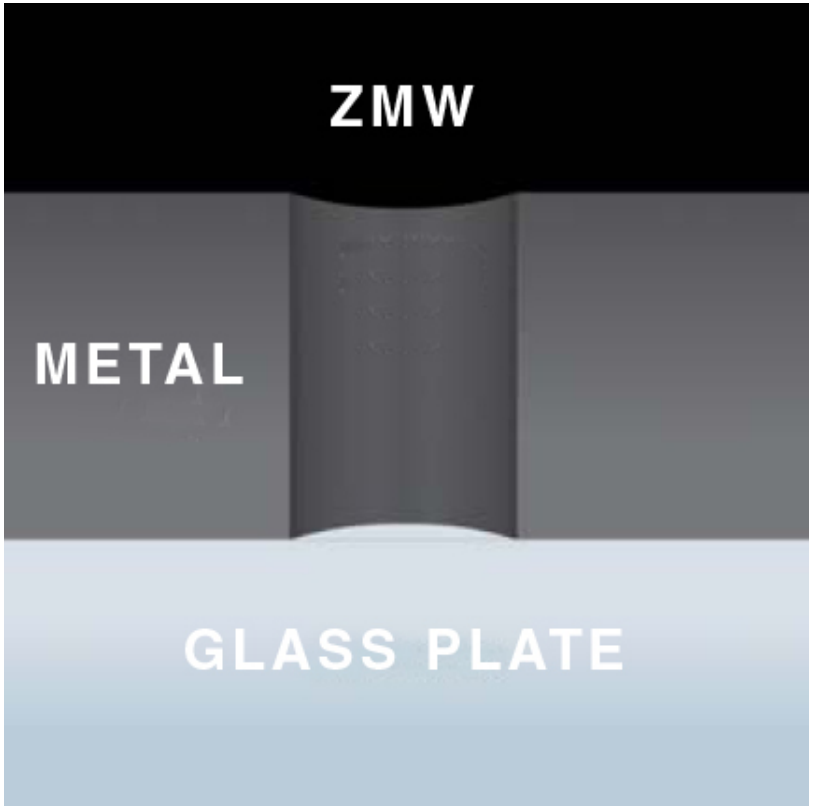
New Technologies

TINY barely captures the size of the PacBio SMRT chip.

Think of the letter O of IN GOD WE TRUST on a US quarter as a tiny bowl.

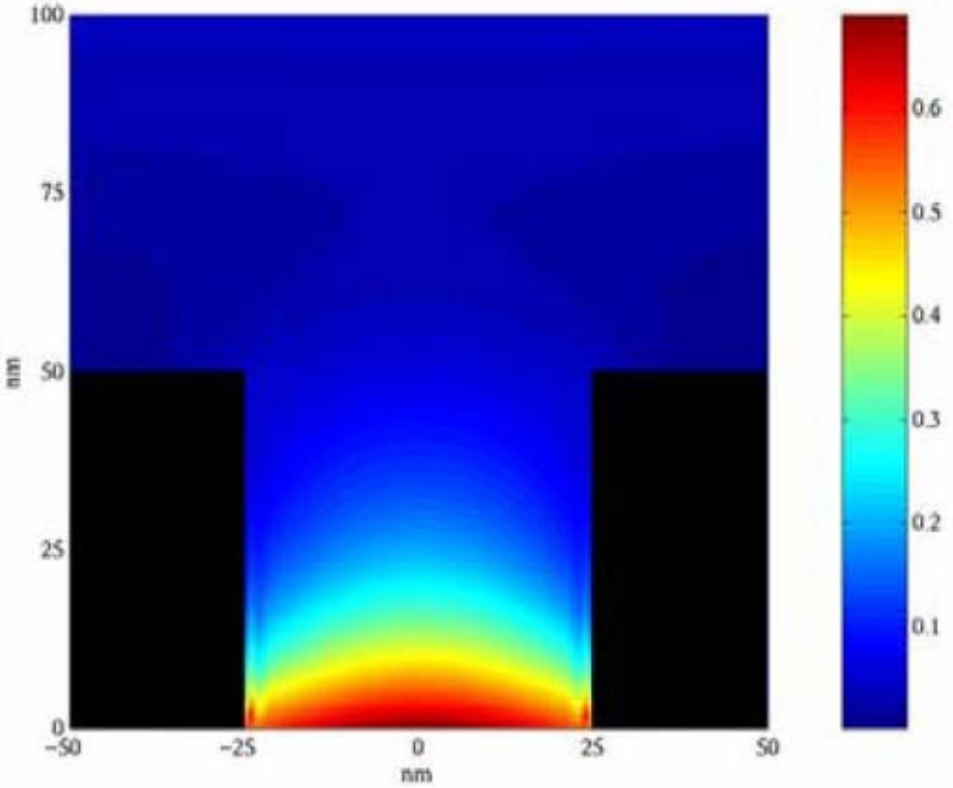
You could pile more than 150 SMRT chips inside that bowl.

New Technologies



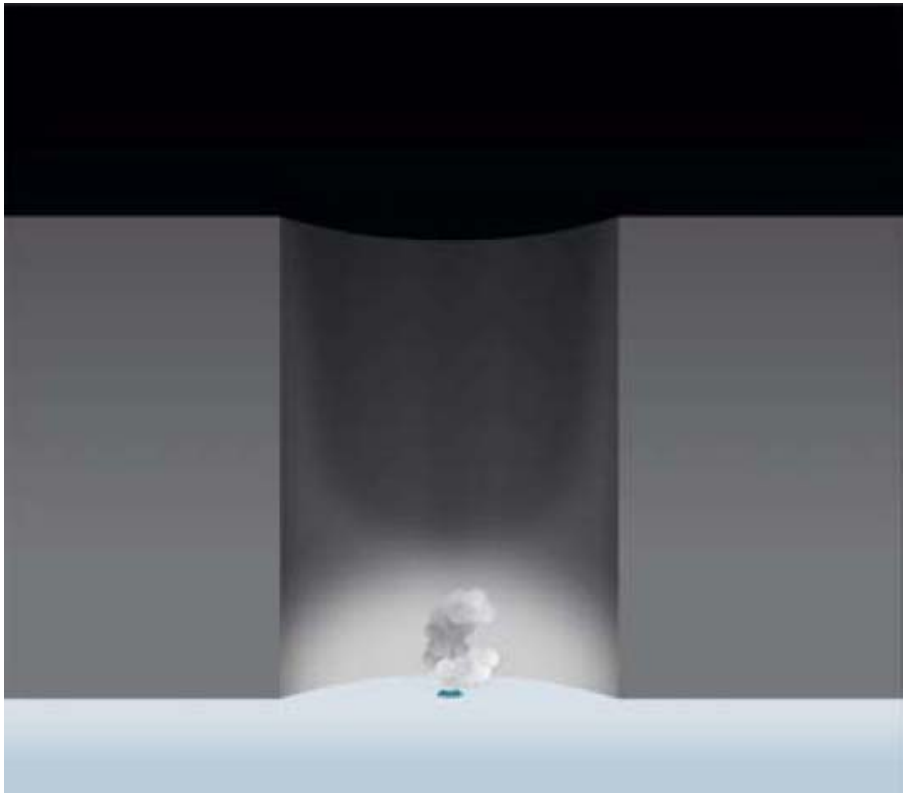
Each zero mode waveguide is a cylindrical hole tens of nanometers in diameter, perforating a thin metal film supported by a transparent glass layer.

New Technologies



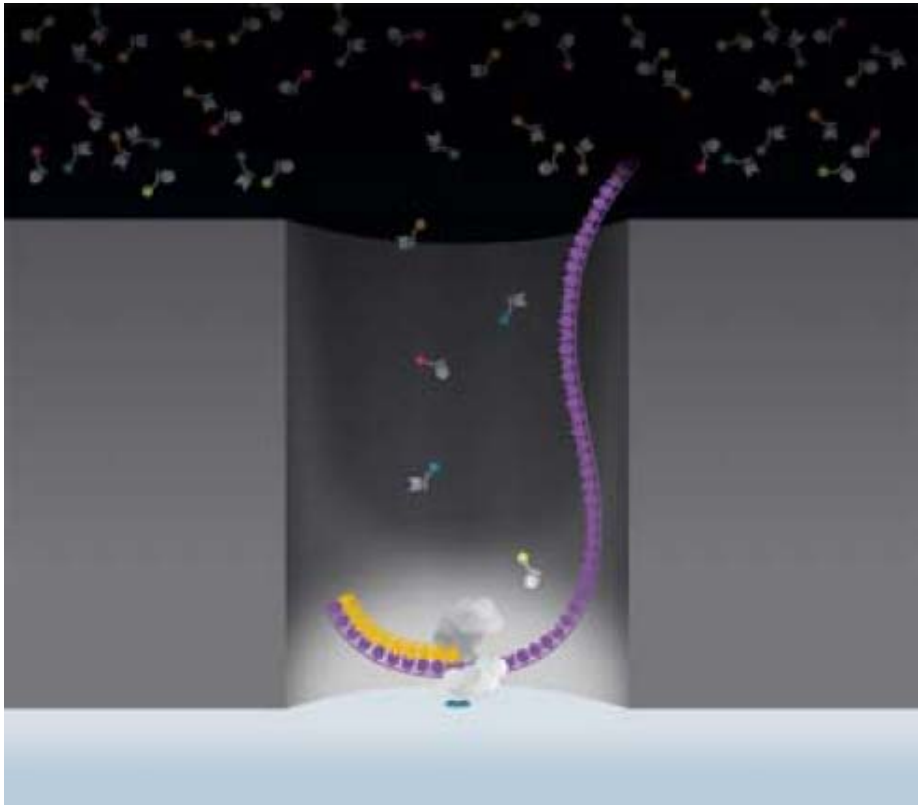
Attenuated light from the excitation beam penetrates only the lower 20 to 30 nm of each waveguide, creating a detection volume of 20 zeptoliters (i.e., 10^{-21} liters).

New Technologies



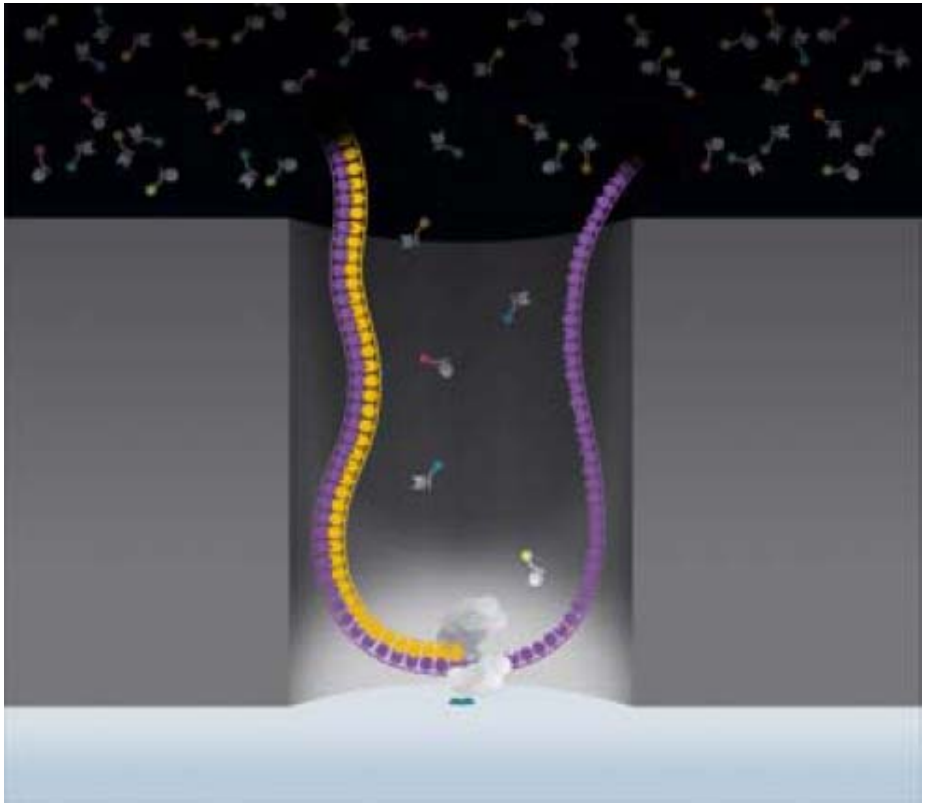
A single DNA polymerase molecule is attached to the bottom of the ZMW using a proprietary-based immobilation process.

New Technologies



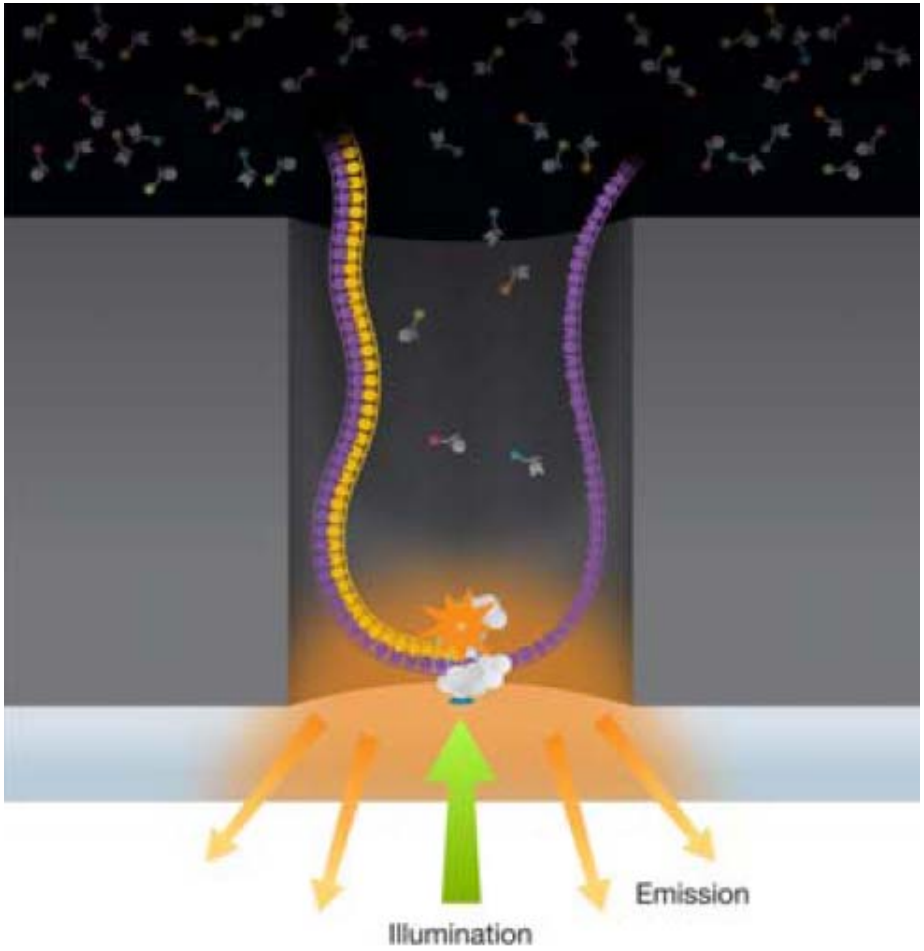
Phospholipid nucleotides are added into the ZMW at the high concentrations required for proper enzyme functioning.

New Technologies



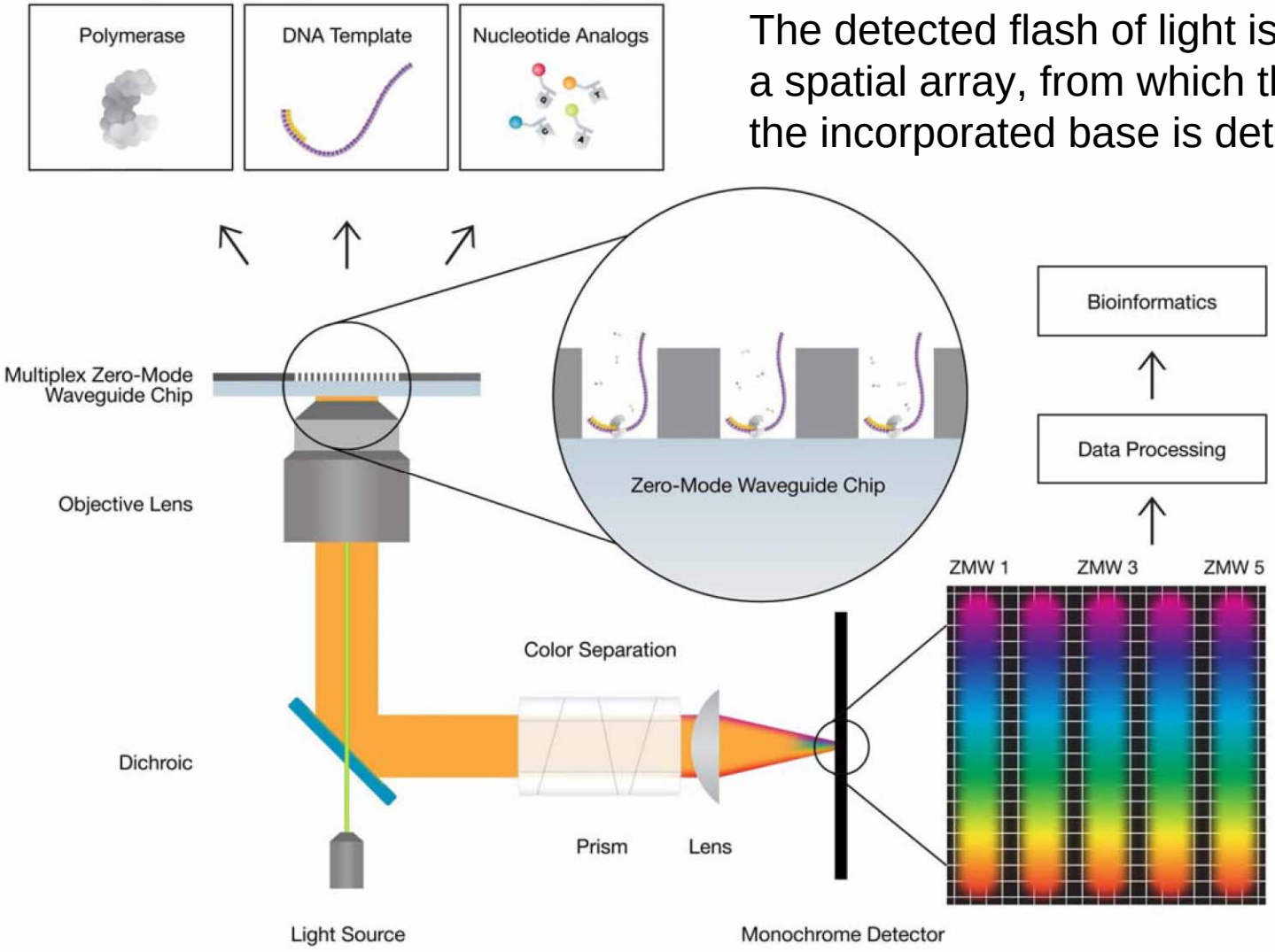
Researchers at PacBio have demonstrated that this approach has the capability to produce reads that are thousands of nucleotides in length.

New Technologies



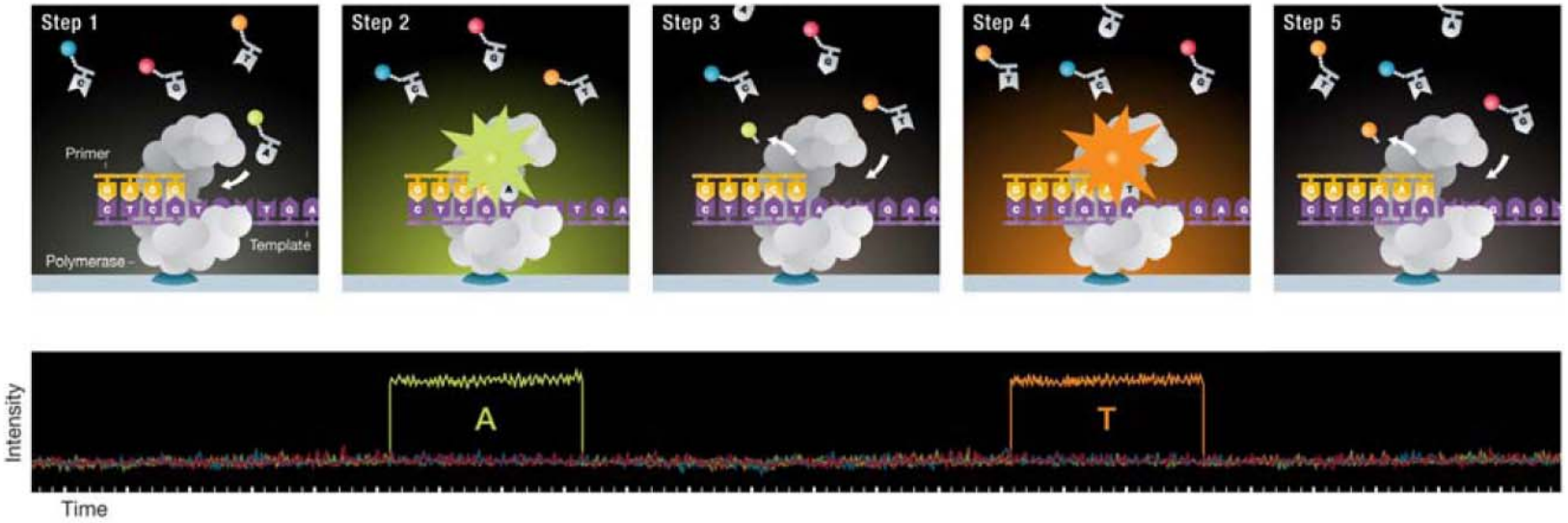
This design supports the continuous and simultaneous excitation and detection of the individual labeled molecules.

New Technologies



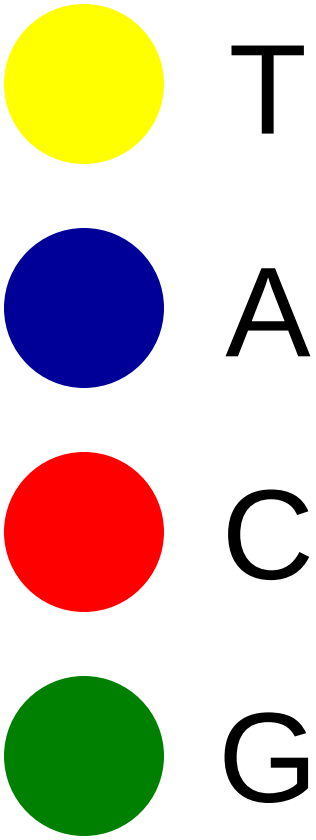
The detected flash of light is separated into a spatial array, from which the identity of the incorporated base is determined.

New Technologies



Sequence data are generated when the enzymatic incorporation of the labeled nucleotide creates a flash of light, which is converted into a base call using optimized algorithms.

New Technologies



Each type of nucleotide is labeled with its own particular color of fluorophore.

In this example, thymine is labeled with yellow, adenine with blue, cytosine with red, and guanine with green.

New Technologies

single well

If we were to look at the bottom of a single well in the Pacific biosciences detection chip, we would see low intensity flickering of different colors of light as different nucleotides rapidly (microseconds) moved into and out of the detection chamber.

Periodically we would see a burst of a sustained (milliseconds) single color, corresponding to the incorporation of a single nucleotide into the growing DNA strand.

By recording the base name corresponding to the particular burst of color, the system can generate, in real time, a single-molecule read of DNA sequence.

New Technologies

single well

If we were to look at the bottom of a single well in the Pacific biosciences detection chip, we would see low intensity flickering of different colors of light as different nucleotides rapidly (microseconds) moved into and out of the detection chamber.

Periodically we would see a burst of a sustained (milliseconds) single color, corresponding to the incorporation of a single nucleotide into the growing DNA strand.

By recording the base name corresponding to the particular burst of color, the system can generate, in real time, a single-molecule read of DNA sequence.

New Technologies

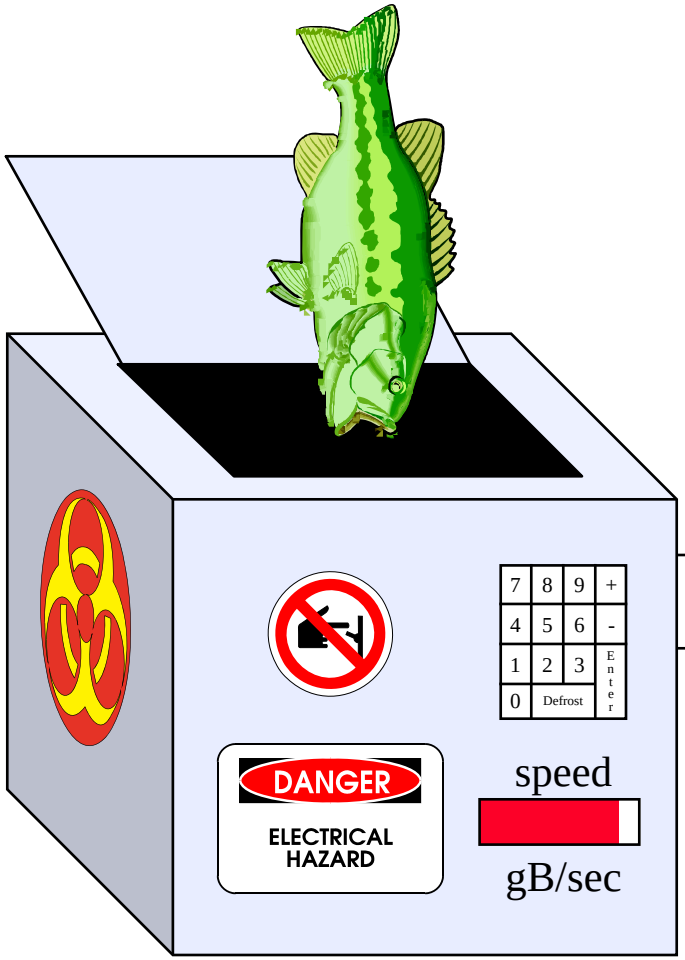
single well

Although this process does run slower than the biological process of DNA synthesis, it is still capable of producing DNA sequence at a truly astounding rate.

In principle, you could start a run on a Pacific Biosciences sequencer before breakfast and by lunchtime have produced as much DNA sequence as was involved in the completion of the human genome project.

Pac BioSciences Sequencer

Fiction becomes reality…

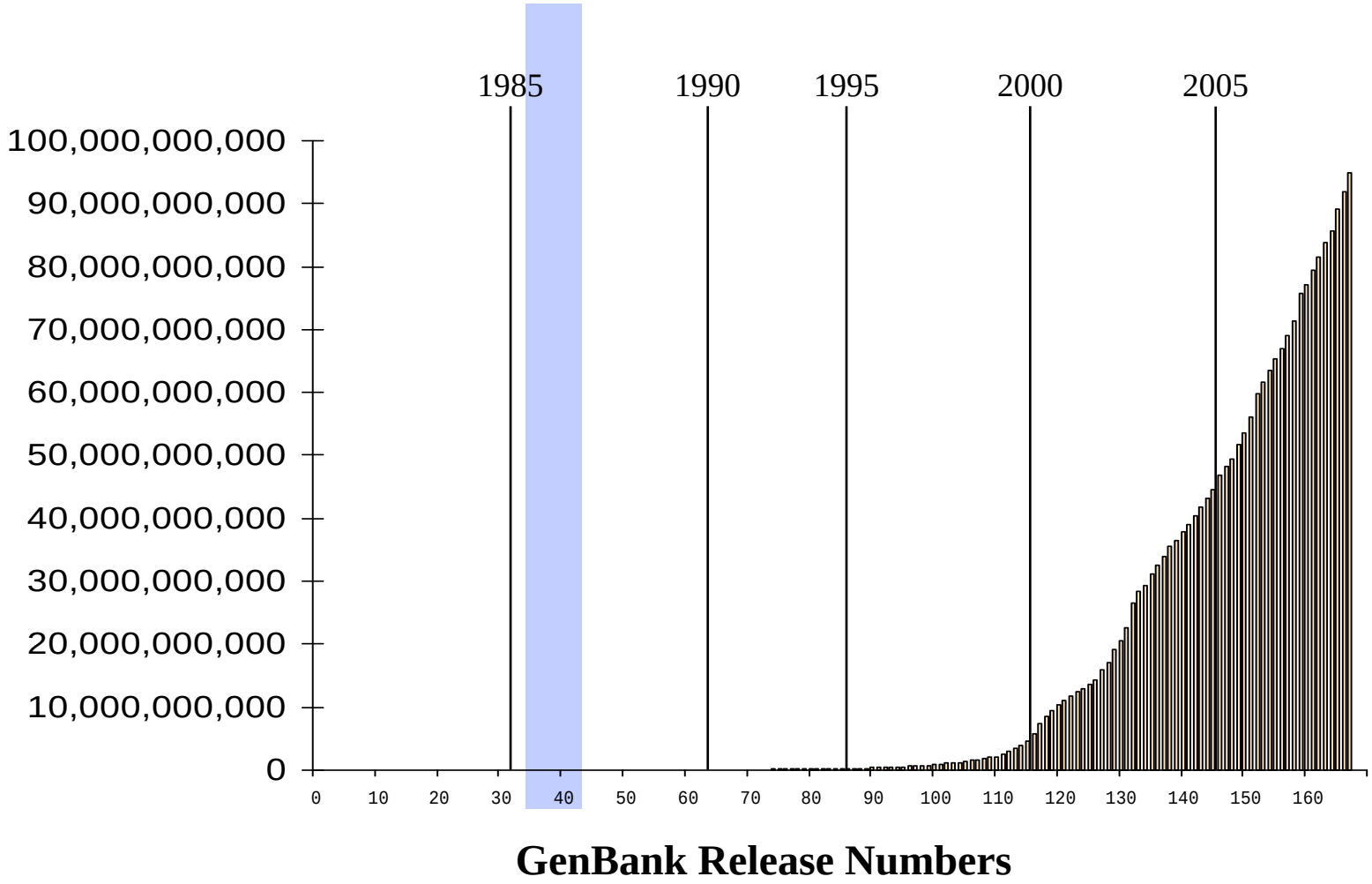


TGCGCATCGCGTATCGATAG

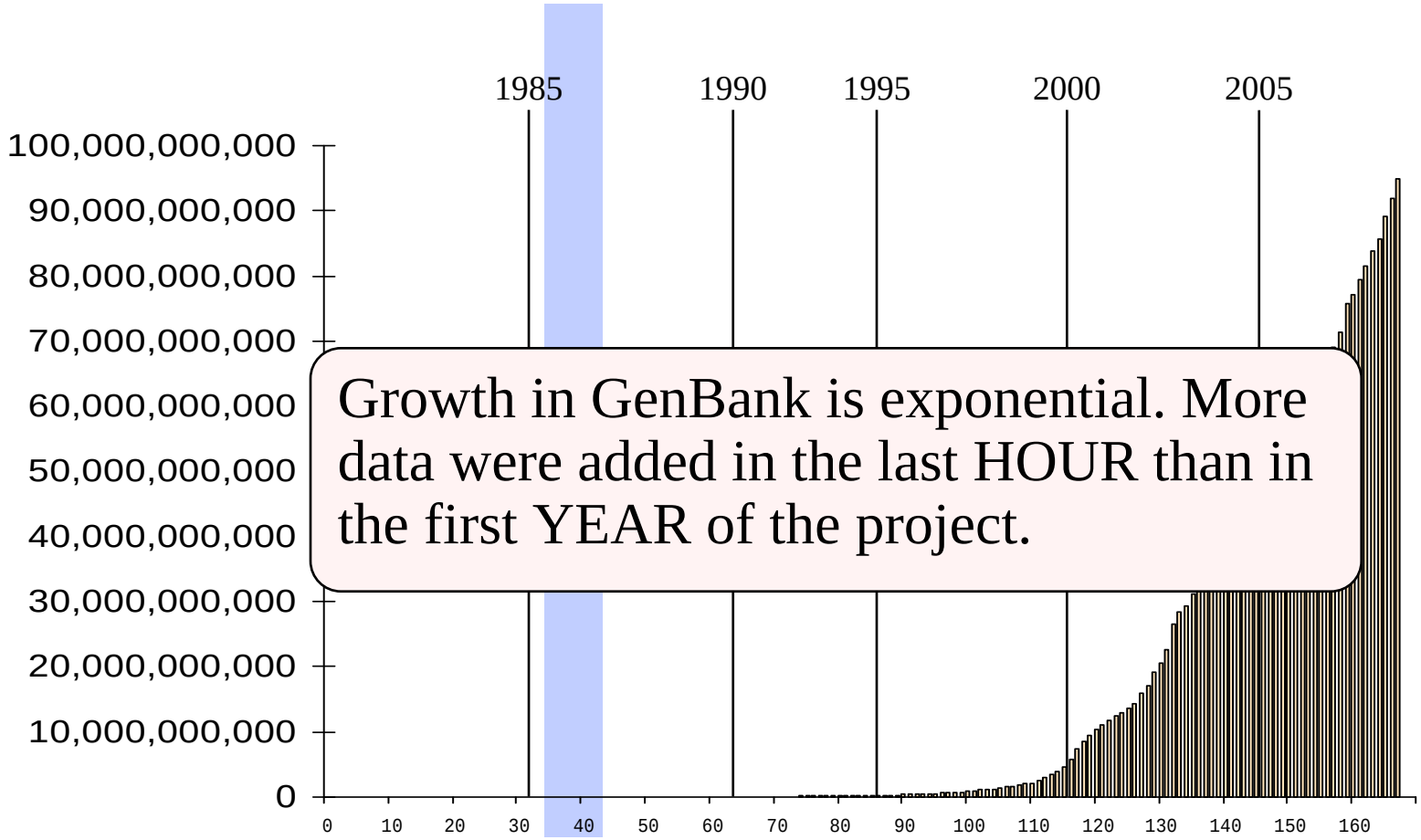
GenBank

Today

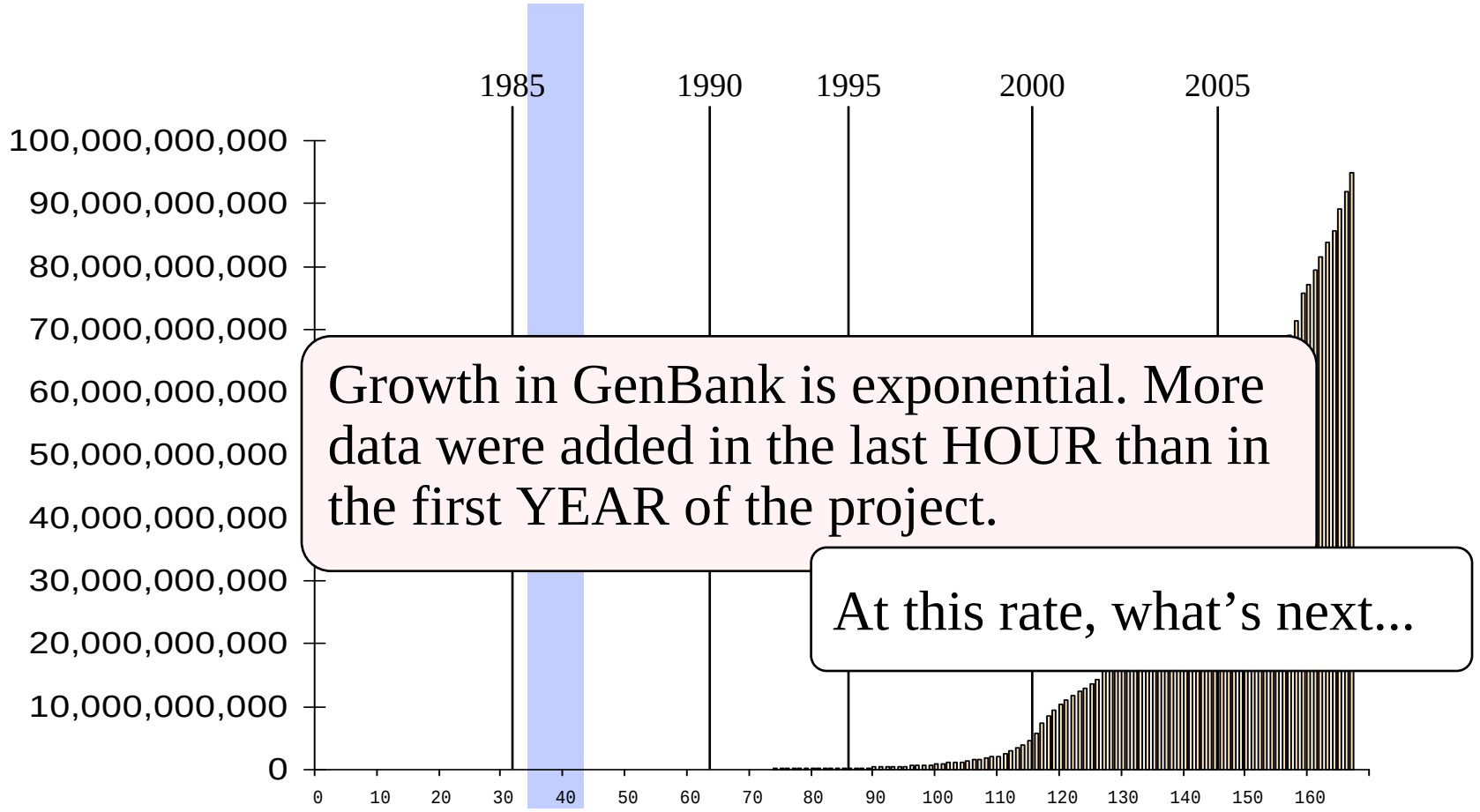
Base Pairs in GenBank (2008)



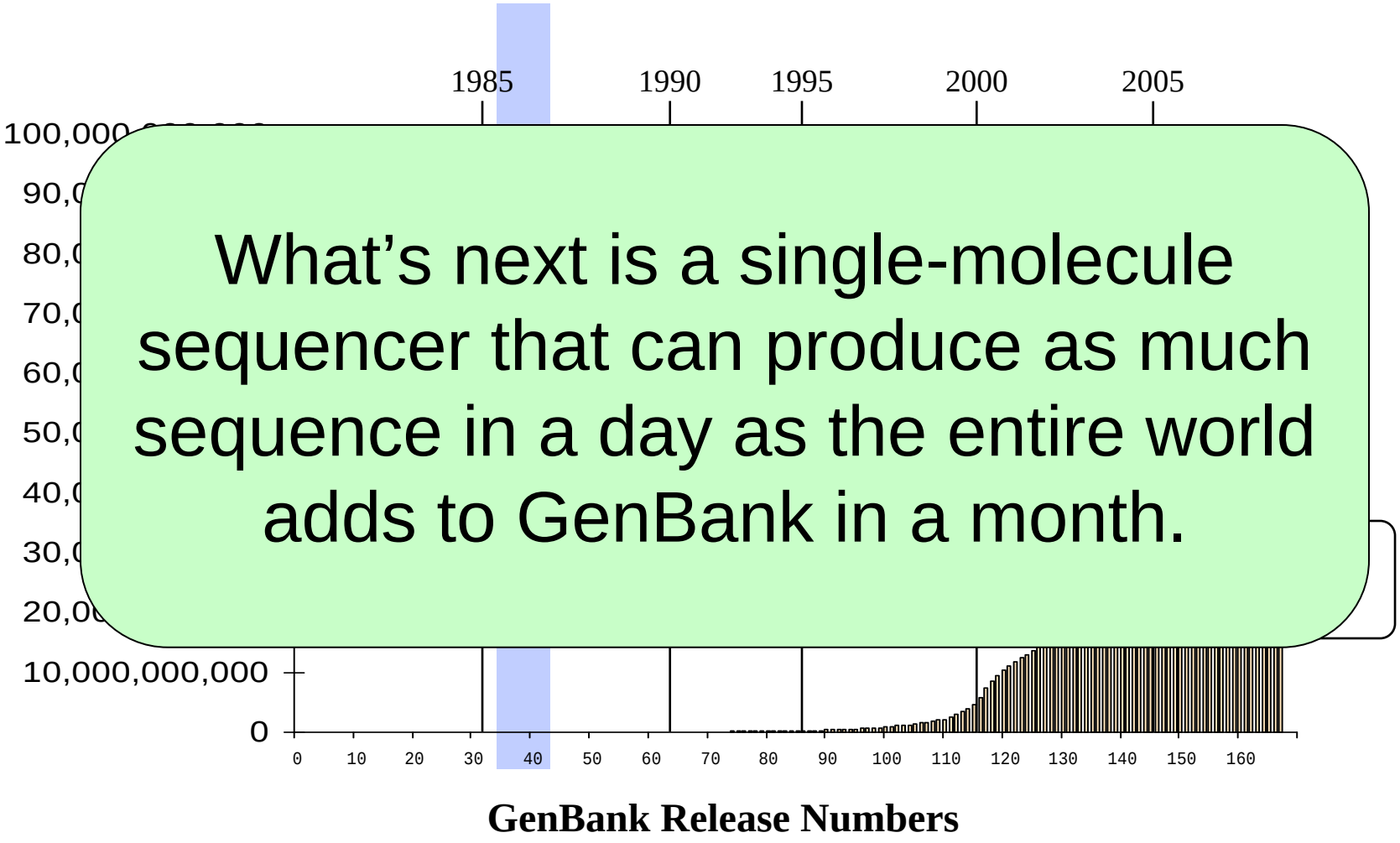
Base Pairs in GenBank (2008)



Base Pairs in GenBank (2008)



Base Pairs in GenBank (2008)



What's next is a single-molecule sequencer that can produce as much sequence in a day as the entire world adds to GenBank in a month.

Problem

Other Challenges

Other Challenges

Many areas need computing horsepower:

- Epidemiological simulation
- Genome wide association studies
- Expression-array analysis
- Proteomics
- Clinical diagnostics
- Image analysis
- …

Summary

- Moore’s Law constantly gives us smaller, better, faster, and cheaper computers. That’s good.

Summary

- Moore's Law constantly gives us smaller, better, faster, and cheaper computers. That's good.
- Hotter, too. That's bad.

Summary

- Moore's Law constantly gives us smaller, better, faster, and cheaper computers. That's good.
- Hotter, too. That's bad.
- Cheaper is also a problem, because people keep coming up with clever ways to use smaller, better, faster, cheaper computers at a rate faster than the computers are getting smaller, better, faster, and cheaper.

Summary

- Moore's Law constantly gives us smaller, better, faster, and cheaper computers. That's good.
- Hotter, too. That's bad.
- Cheaper is also a problem, because people keep coming up with clever ways to use smaller, better, faster, cheaper computers at a rate faster than the computers are getting smaller, better, faster, and cheaper.
- As a result, our data centers are filling up.

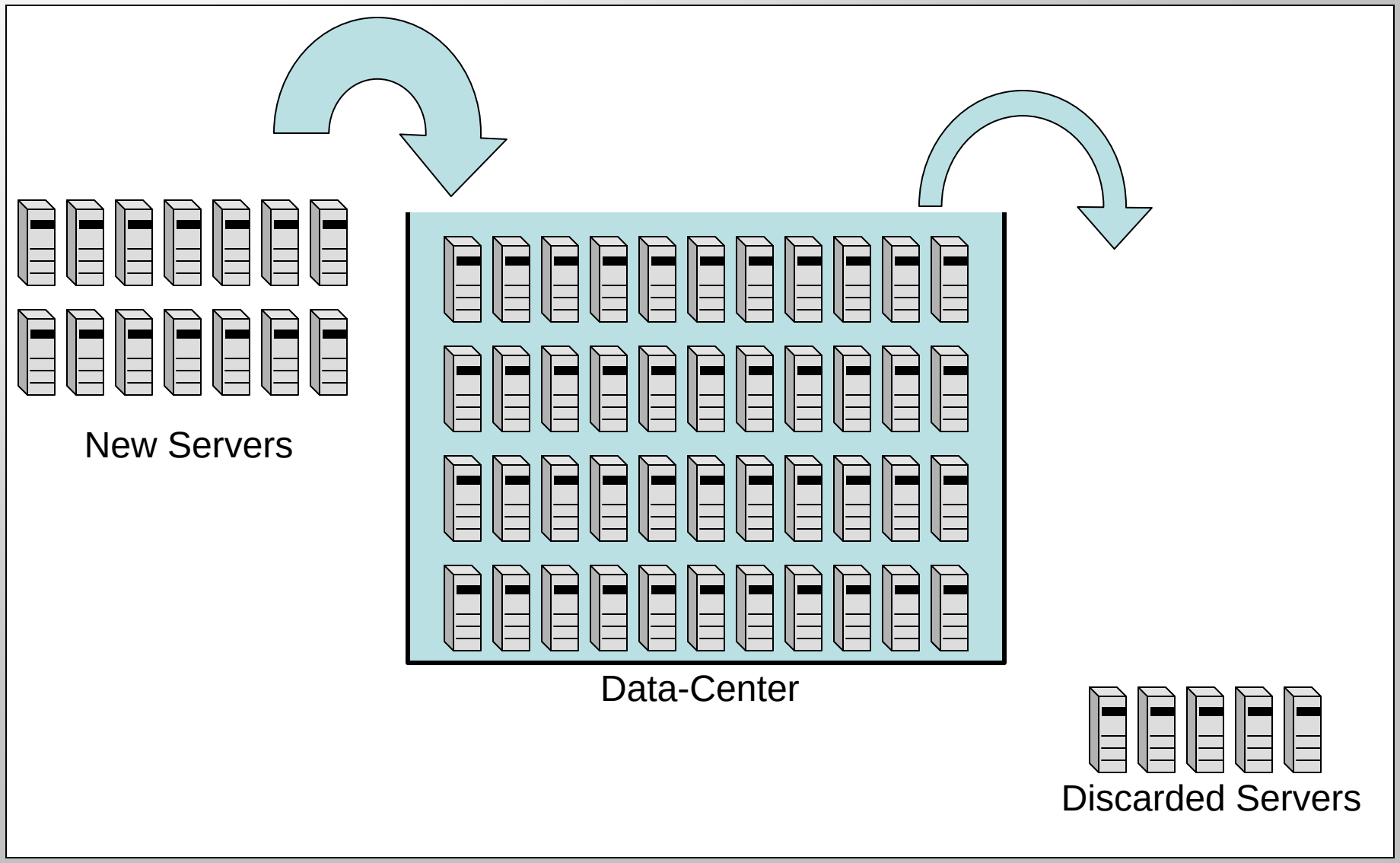
Solutions

Summary

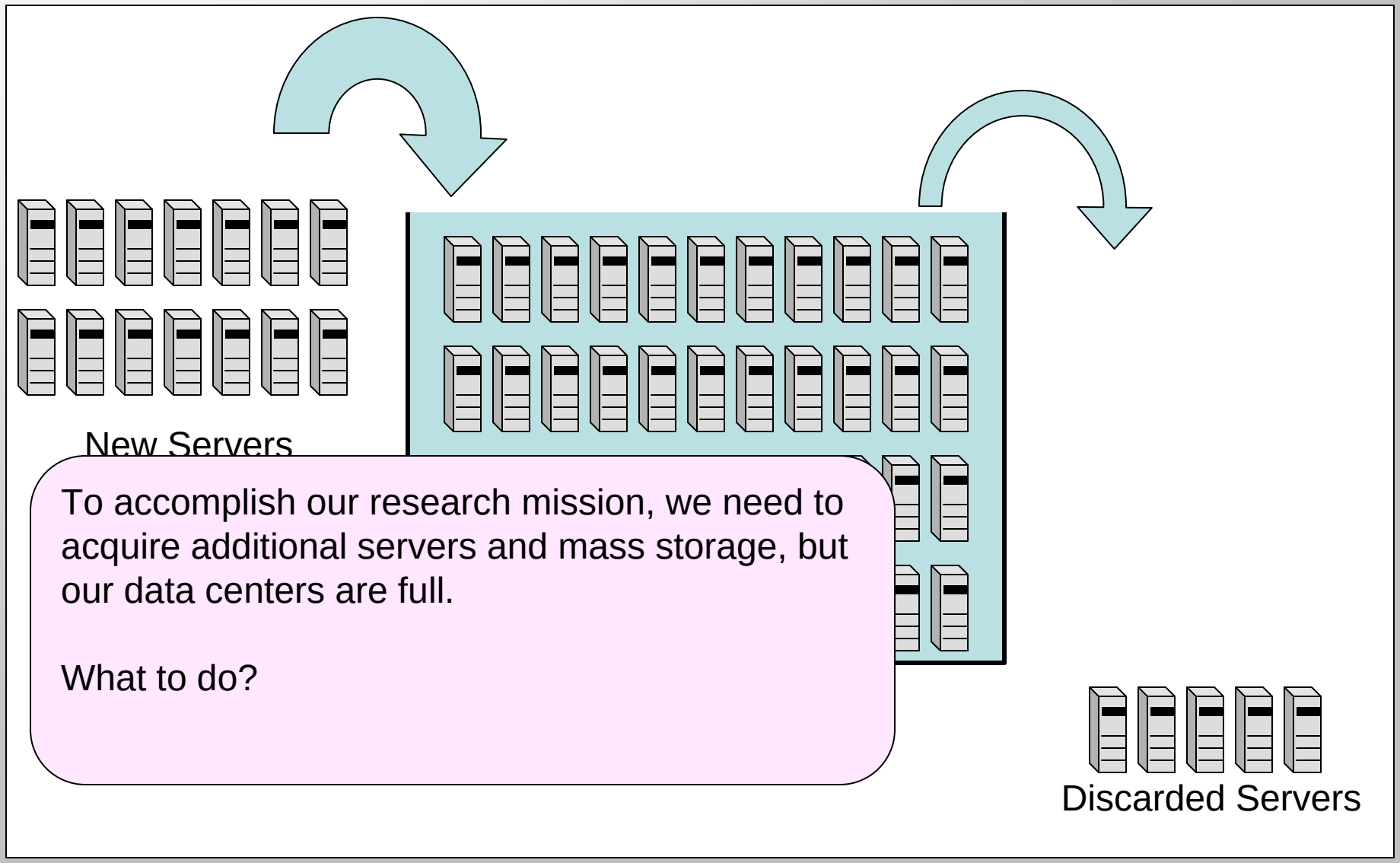
Build More Data-Center Space

- Sounds easy enough:
  - Dedicate or acquire space
  - Build a data center
- Challenges exist:
  - Current technology places great demands on data center design
  - Need lots of power
  - And lots of cooling
  - This can be very expensive

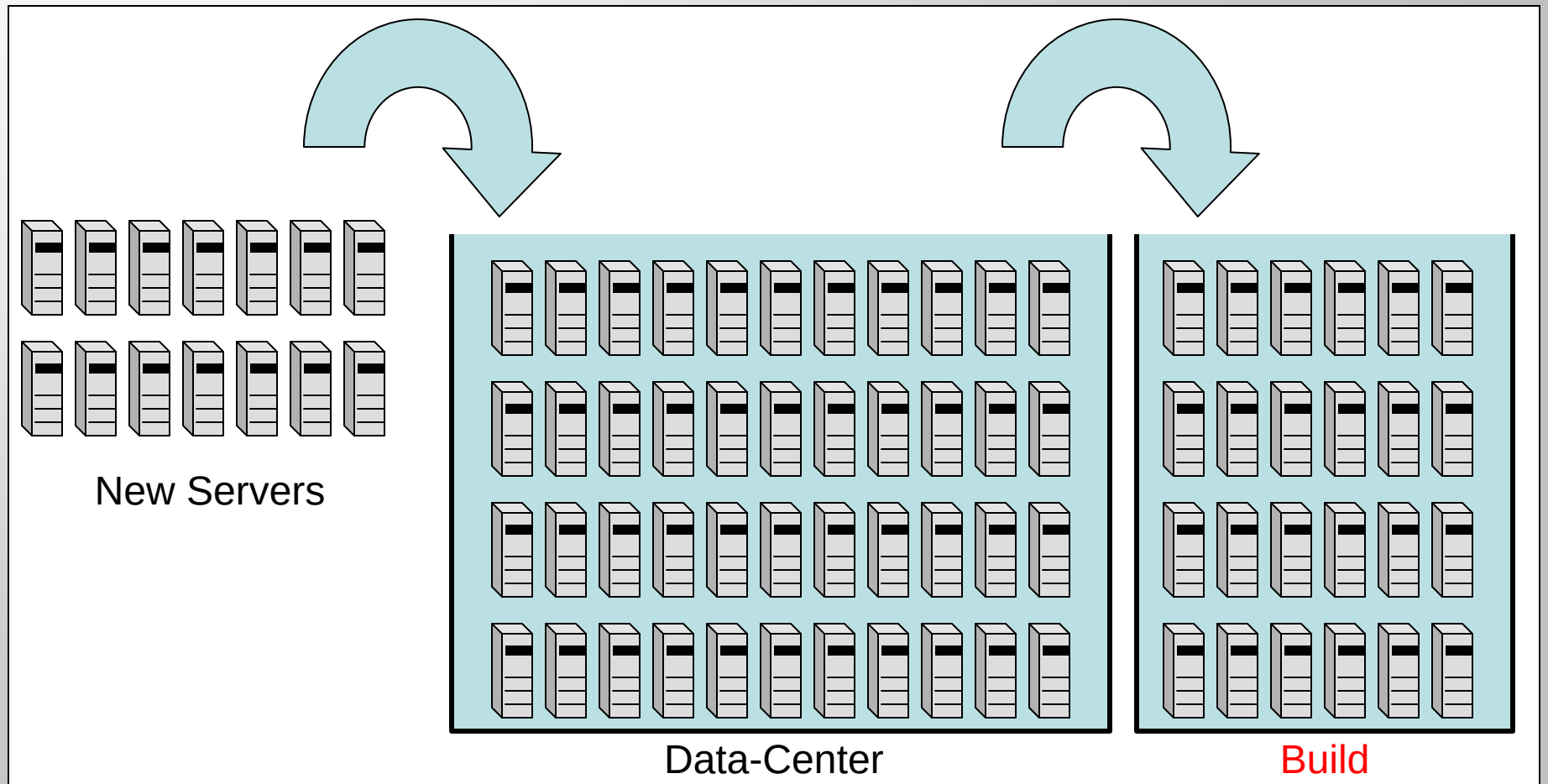
Data-Center problem



Data-Center problem

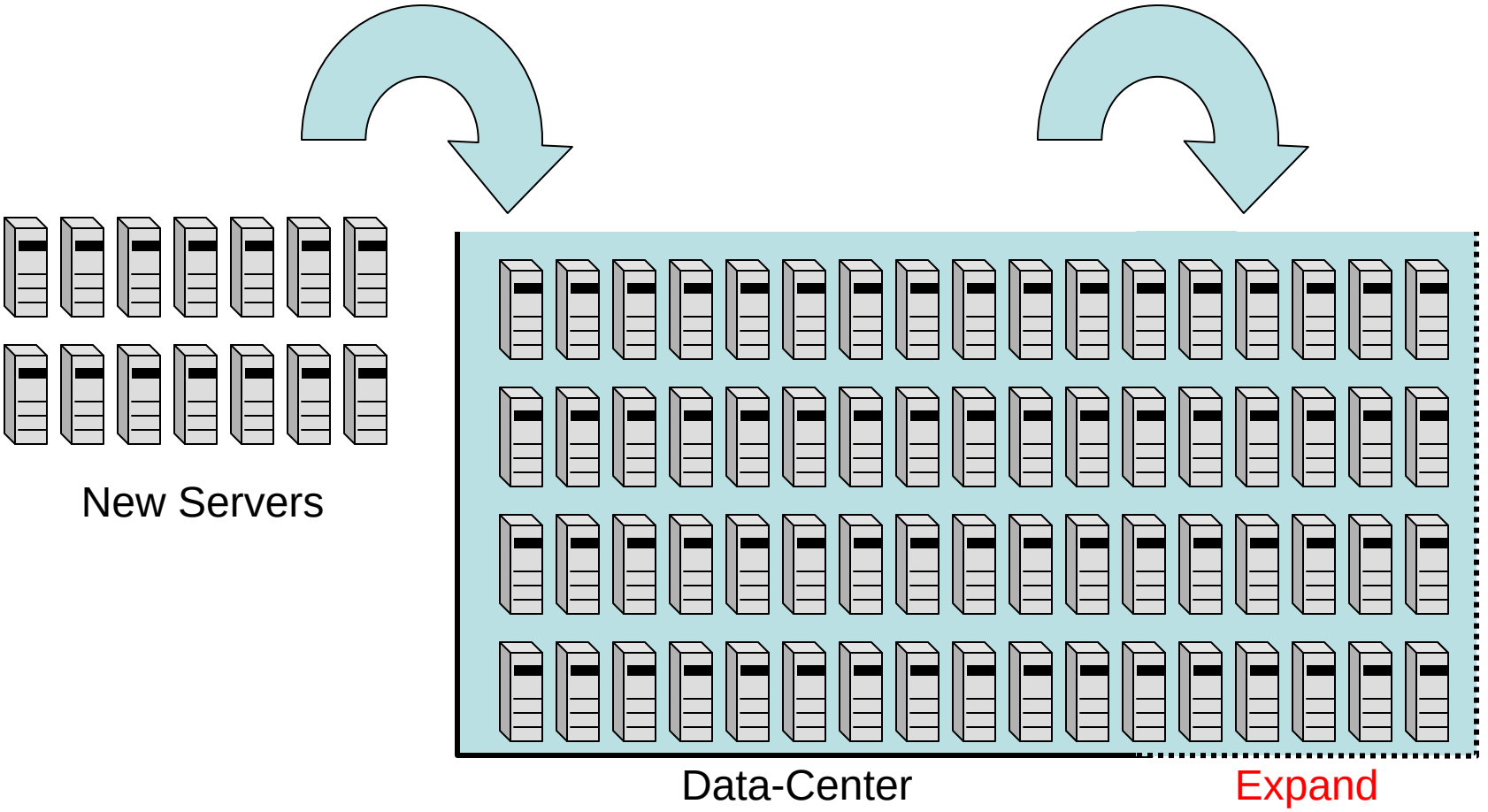


Data-Center Options



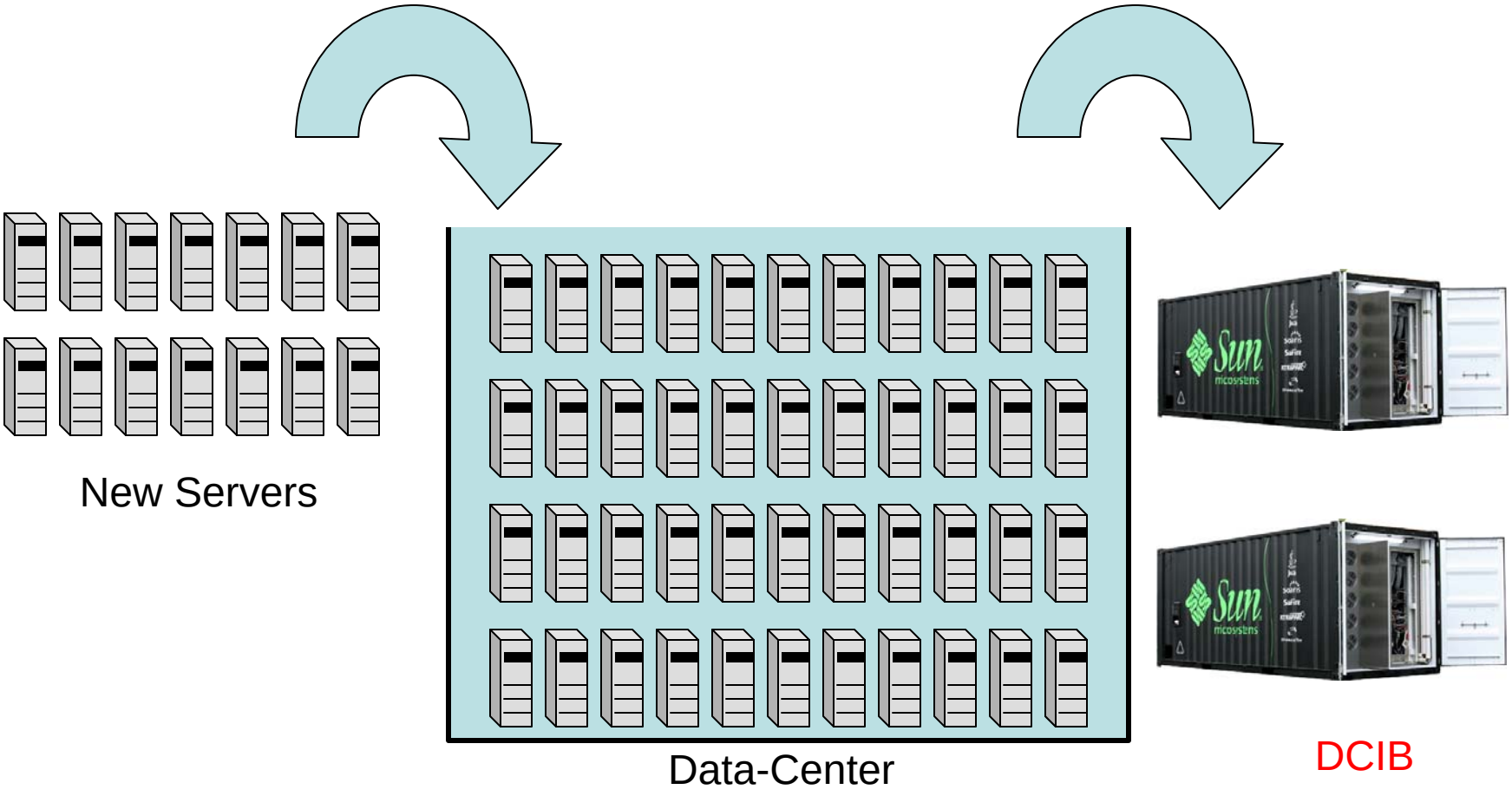
Build New Data Center on Campus

Data-Center Options



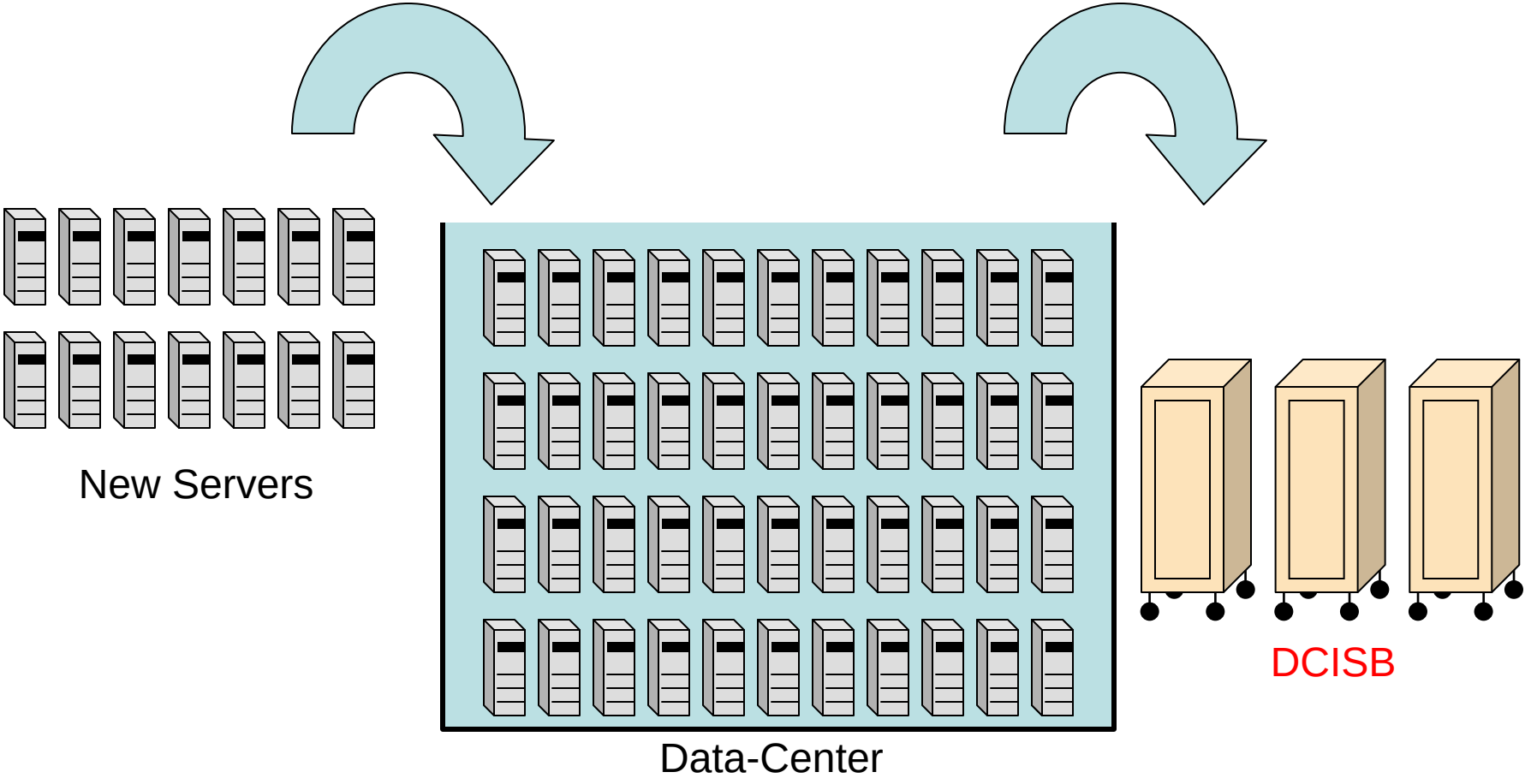
Expand Existing Data Center(s)

Data-Center Options



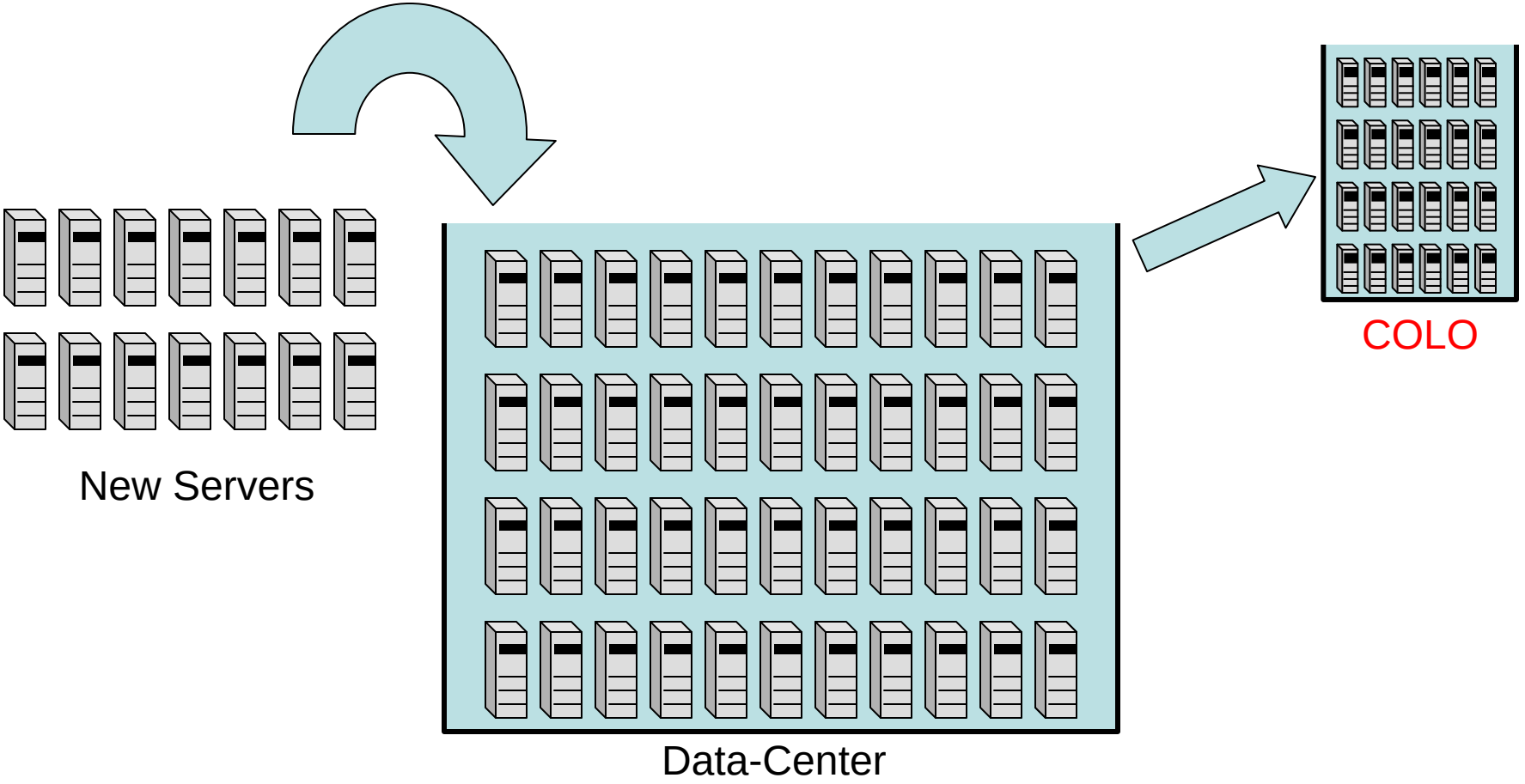
Install Prefabricated Data Center(s) in a Box

Data-Center Options



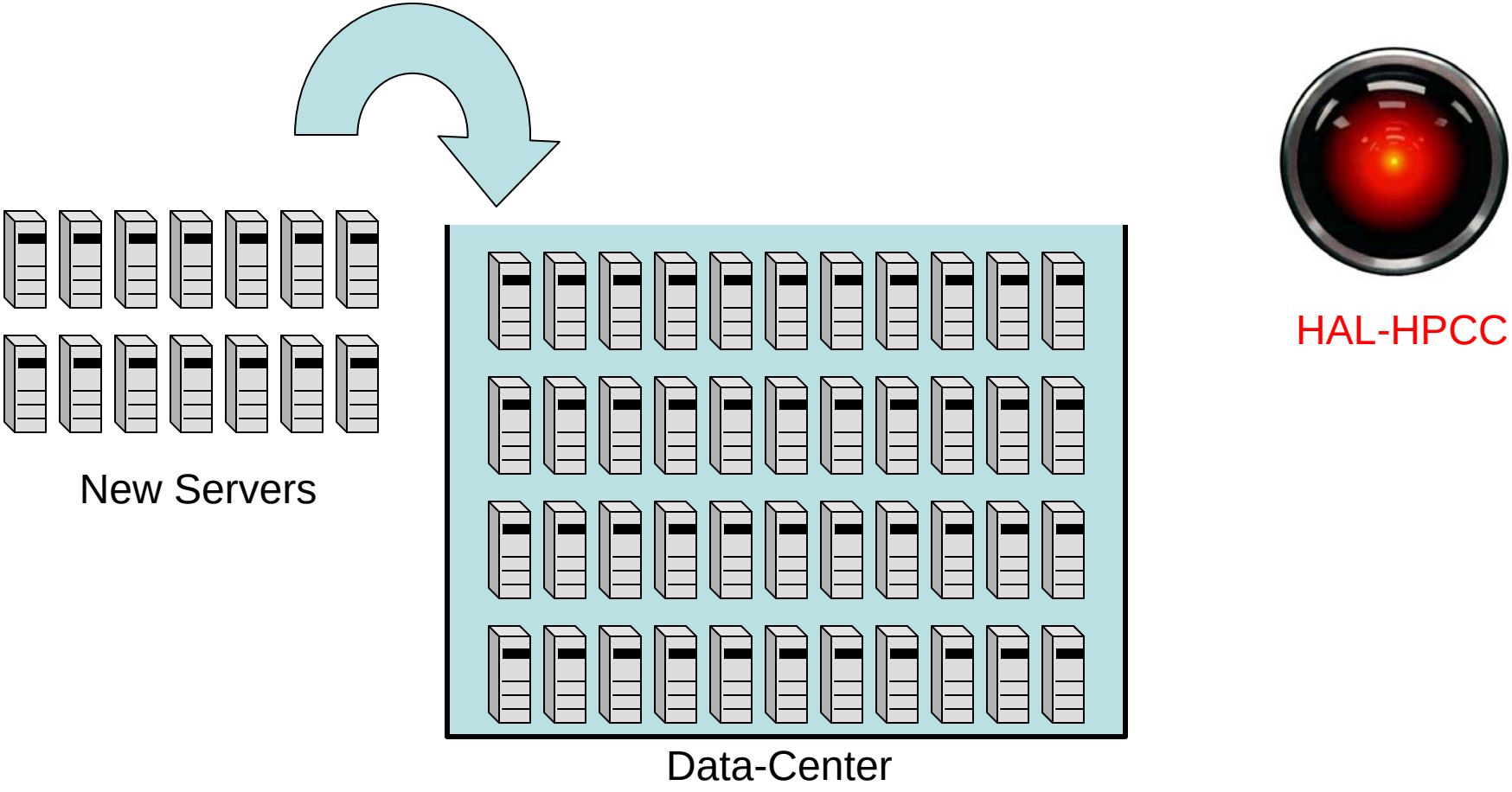
Use Mobile, Self-Contained Cabinets in Unoccupied Lab Space

Data-Center Options



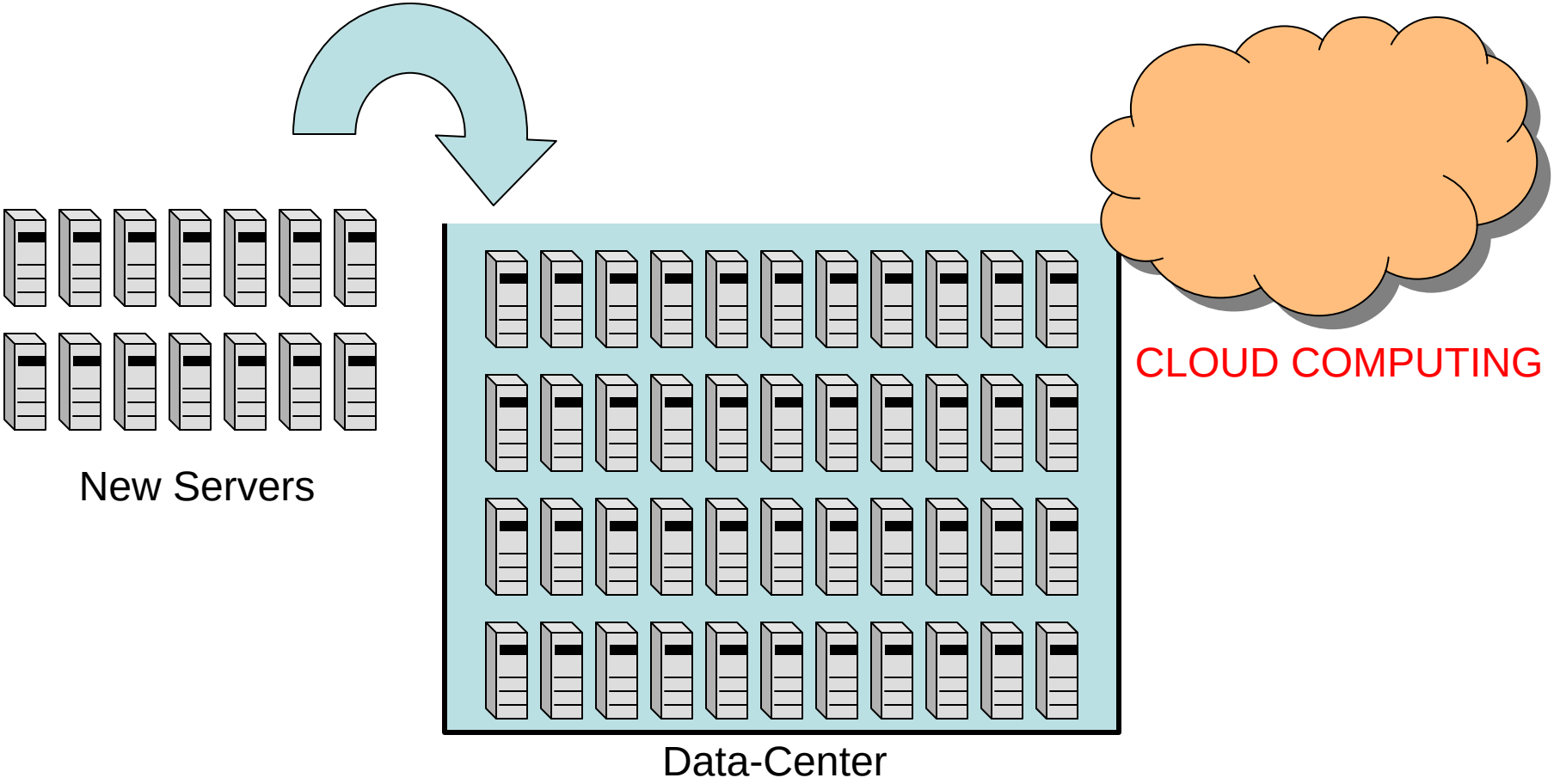
Use an Off-site Co-Location Facility

Data-Center Options



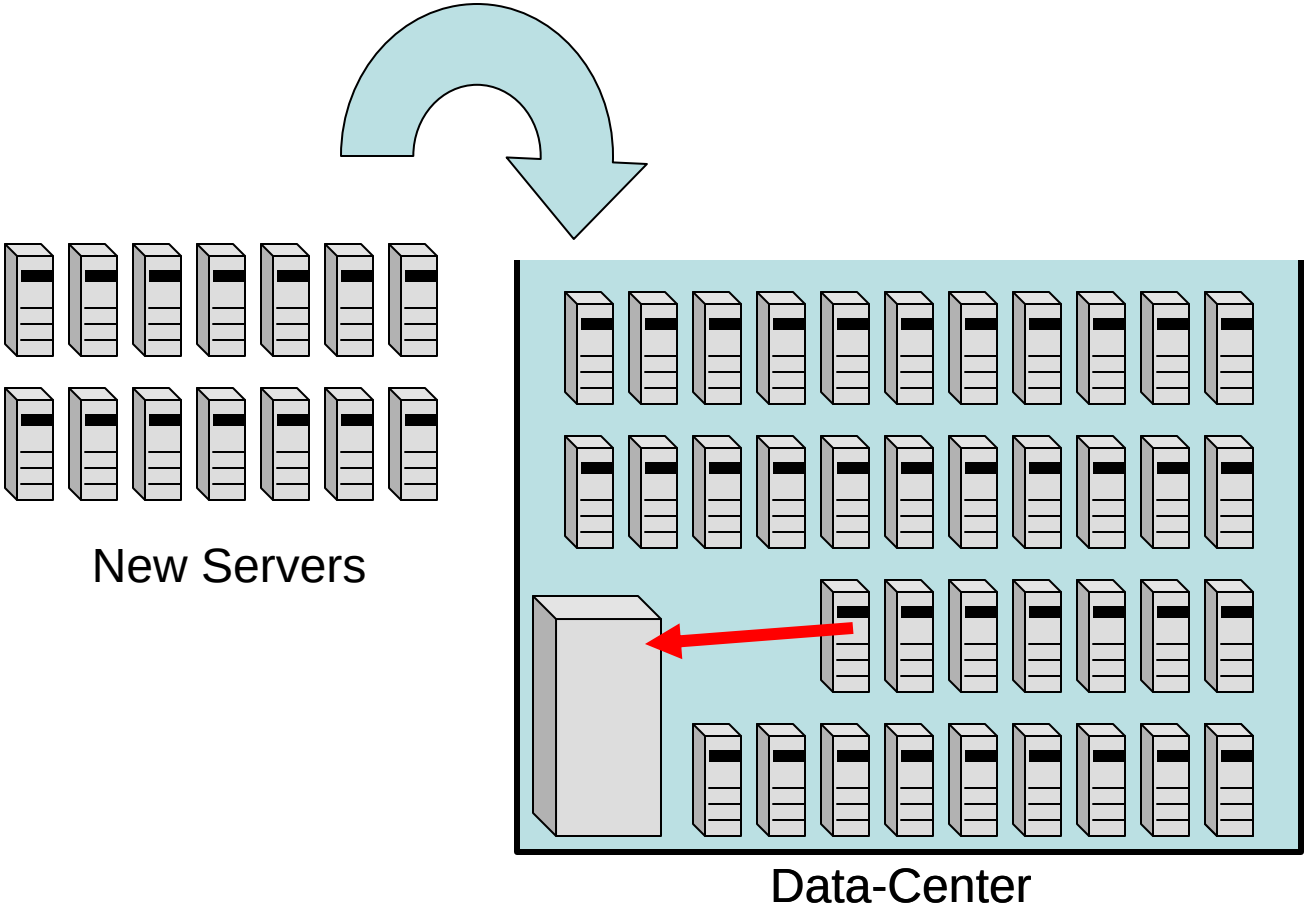
Outsource Specialized High-Performance Computing

Data-Center Options



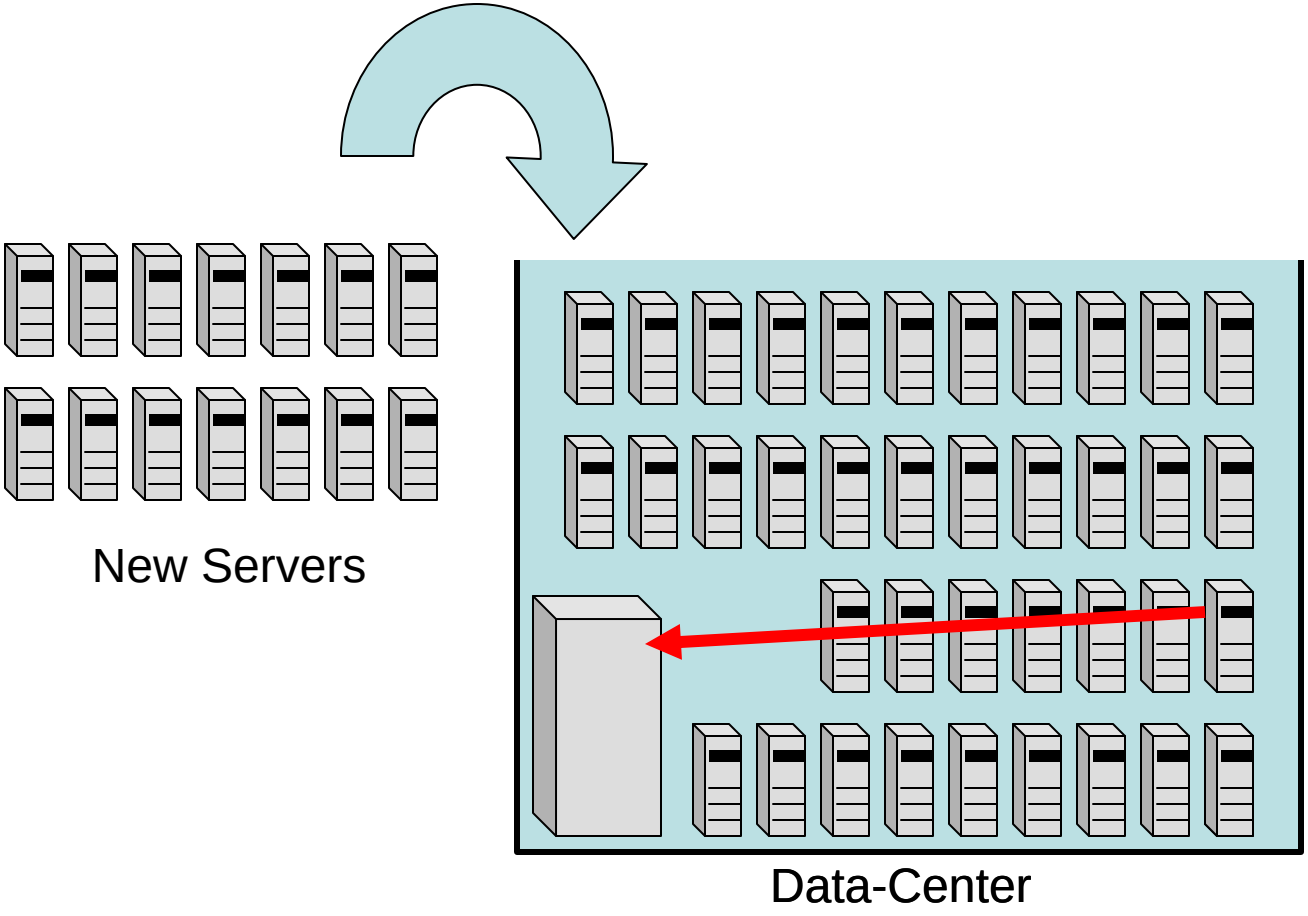
Use the Cloud

Data-Center Options



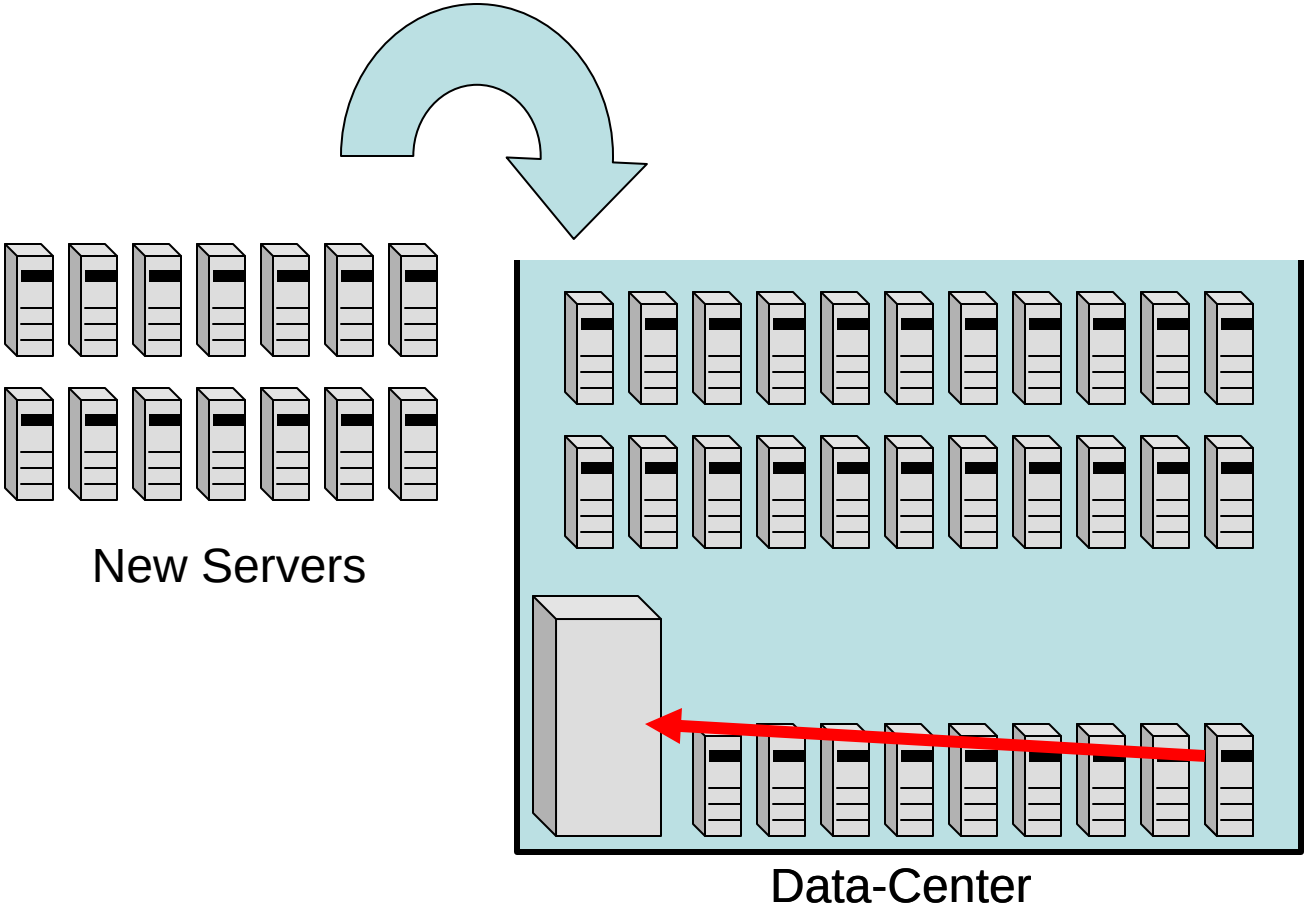
Virtualization: Put Servers Inside Other Servers

Data-Center Options



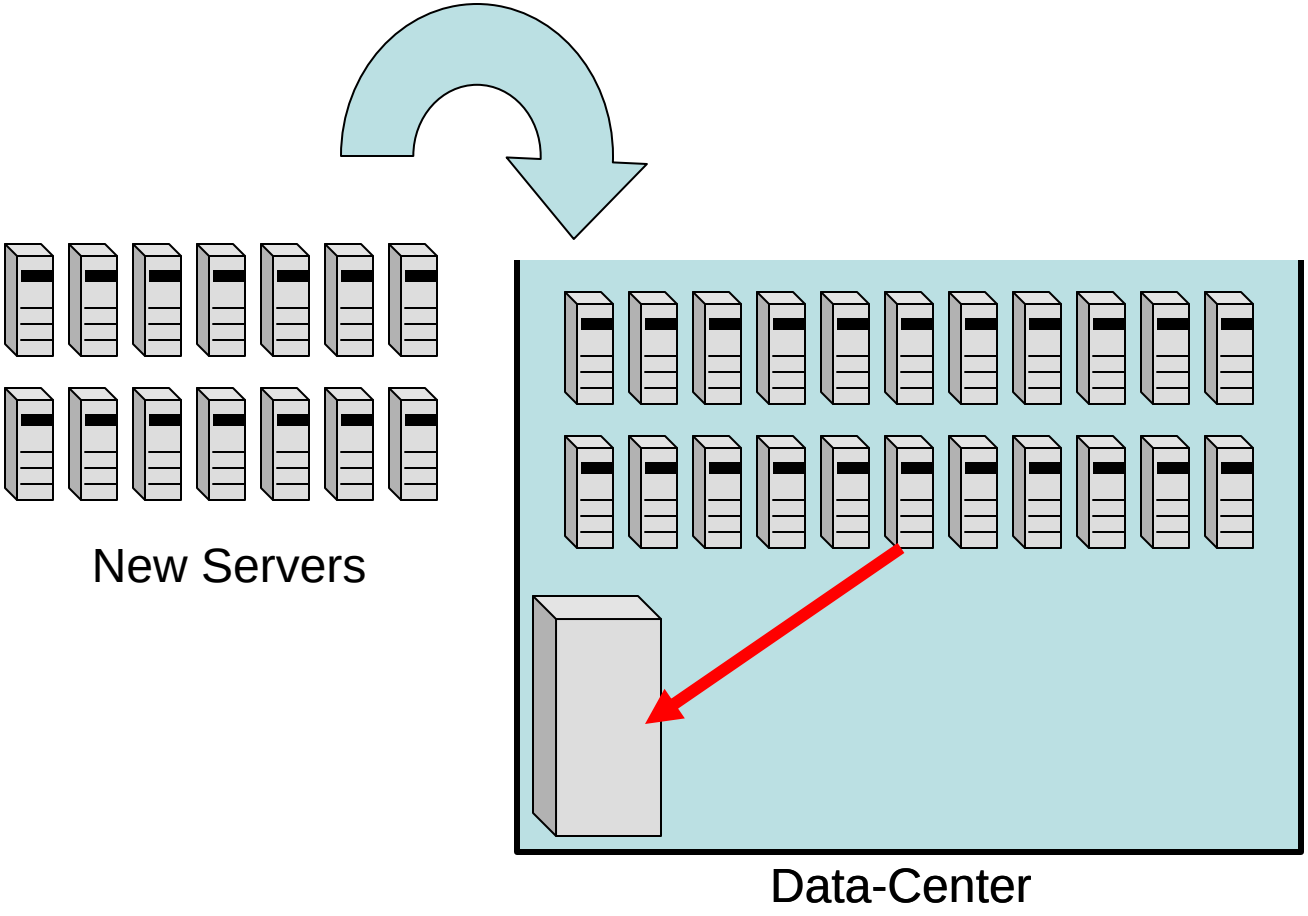
Virtualization: Put Servers Inside Other Servers

Data-Center Options



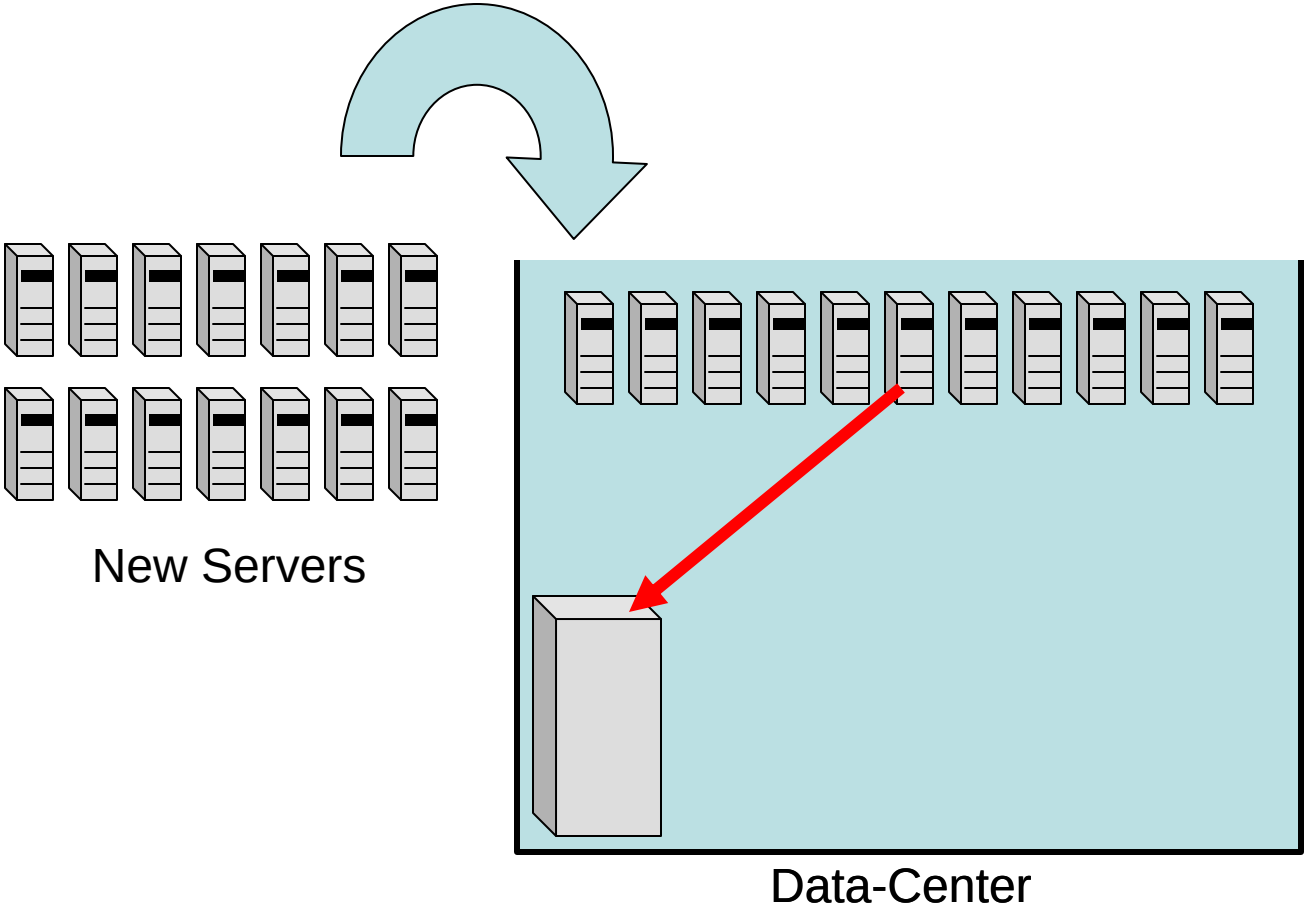
Virtualization: Put Servers Inside Other Servers

Data-Center Options



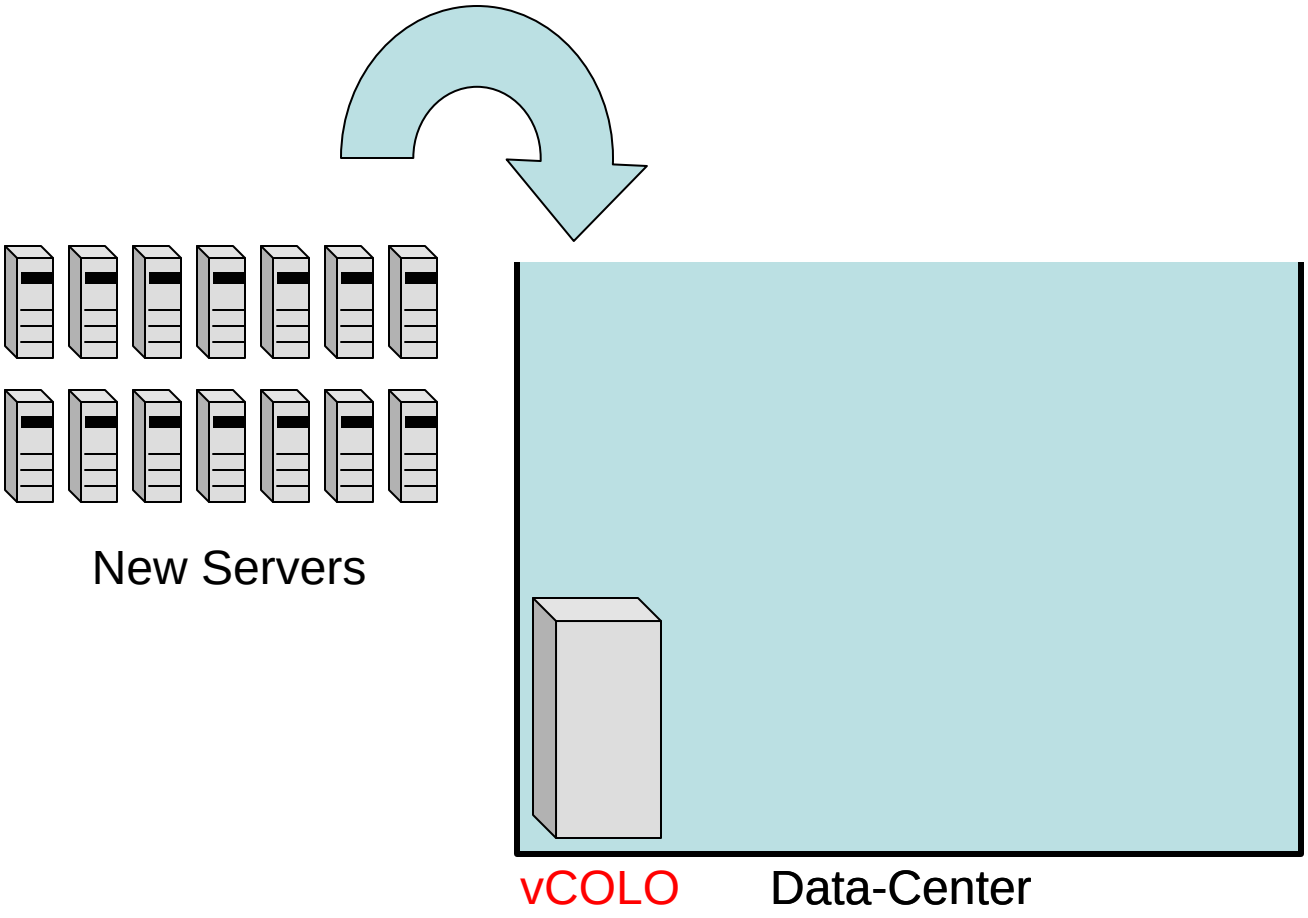
Virtualization: Put Servers Inside Other Servers

Data-Center Options



Virtualization: Put Servers Inside Other Servers

Data-Center Options



Virtualization: Put Servers Inside Other Servers

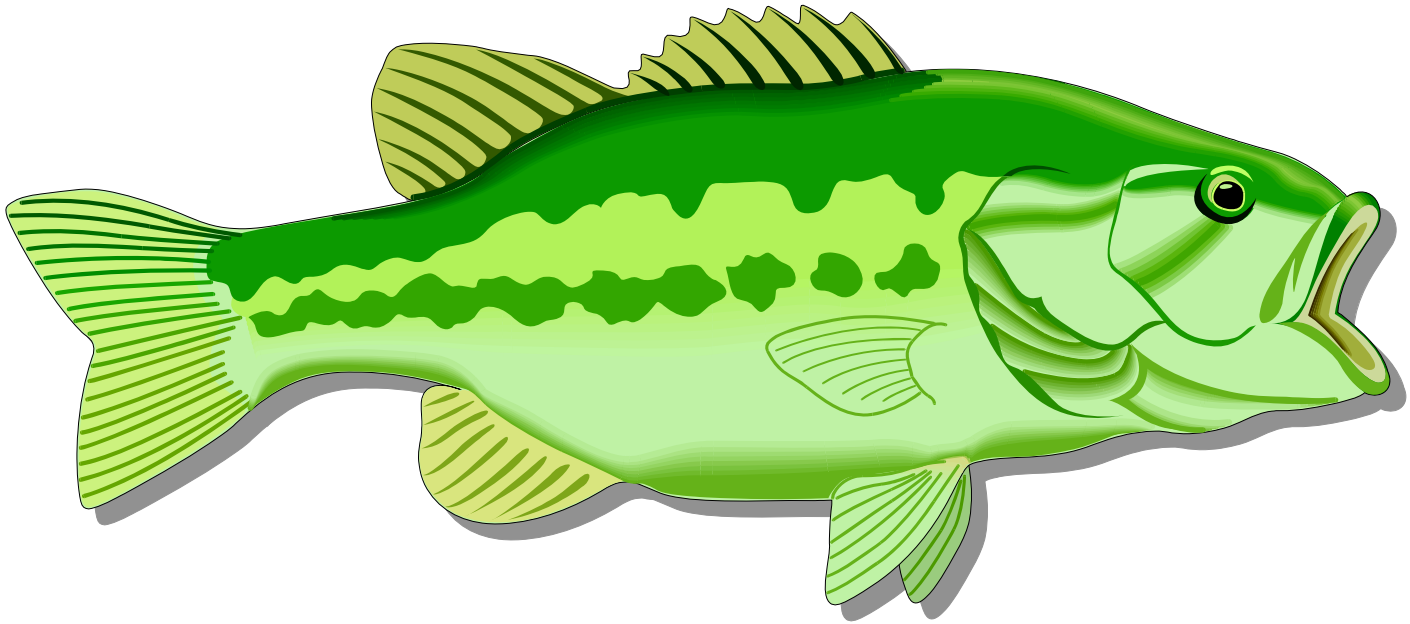
Decisions Are Tricky

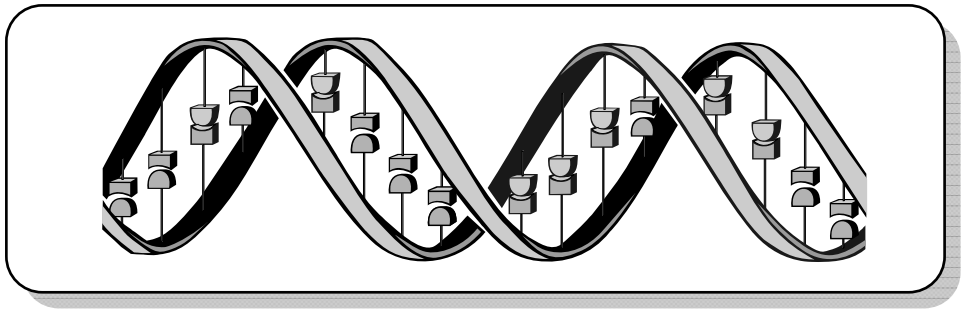
- The options have different advantages and disadvantages.
- The options have different long- and short-term cost profiles.
- Some are primarily expense, others capital. Some are both.
- The options have different useful life expectancies.
- Comparing these is less like comparing apples and oranges and more like comparing apples and Buicks.

END

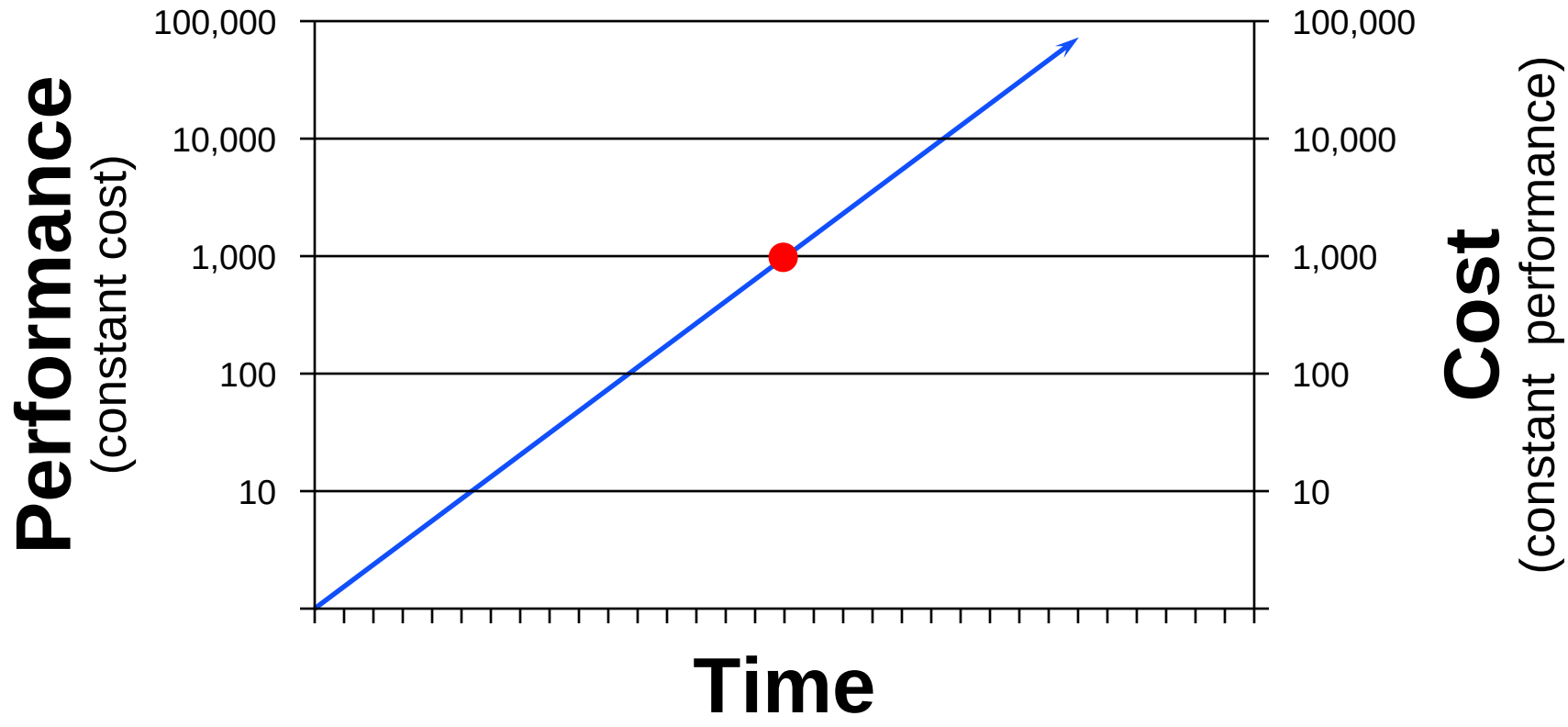
EXTRAS

Samples to Sequence



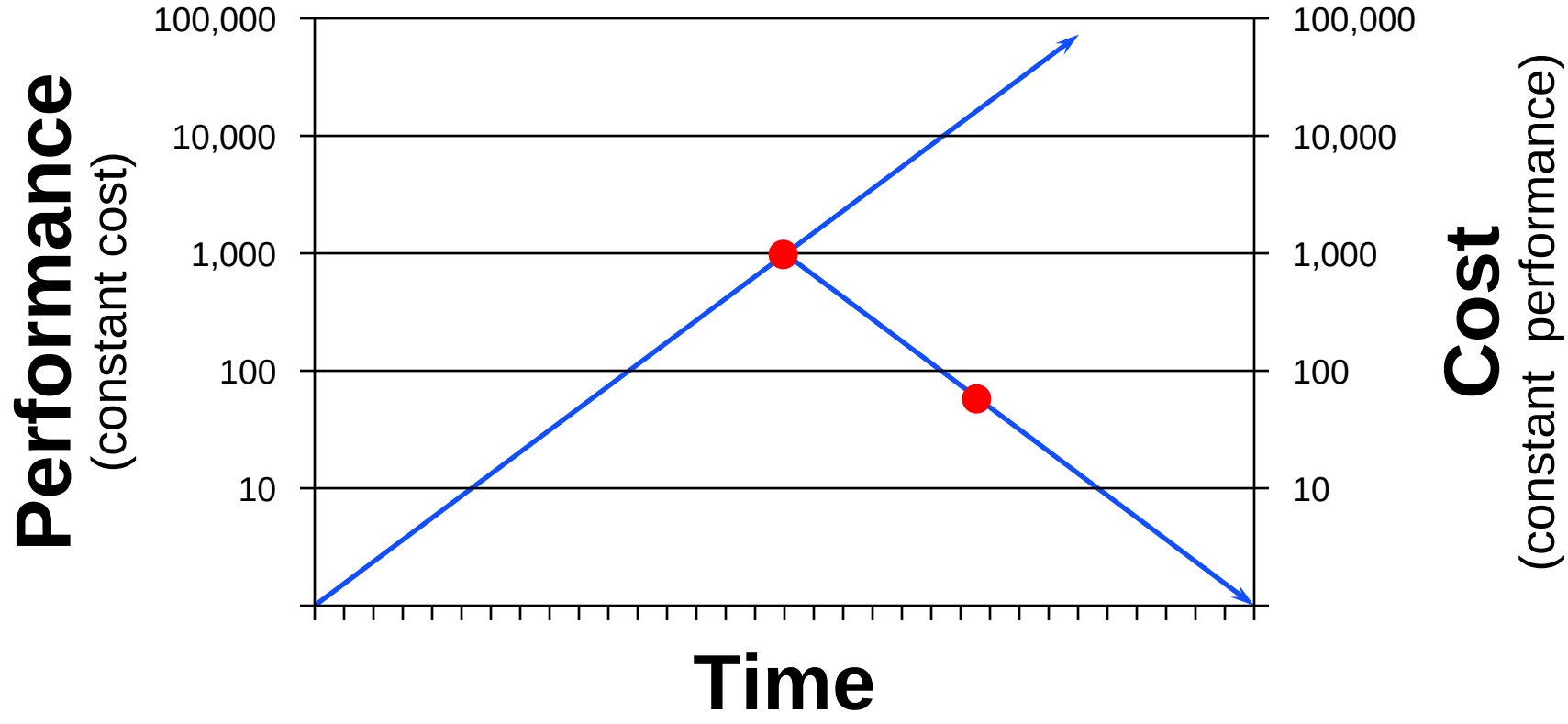


Moore's Law



Performance at constant cost increases exponentially.

Moore's Law



Cost at constant performance decreases exponentially.