Artificial intelligence is changing the world

<table>
<thead>
<tr>
<th>Today</th>
<th>By 2020</th>
<th>By 2020</th>
<th>By 2020</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;30,000</td>
<td>85%</td>
<td>$47B</td>
<td>20%</td>
</tr>
</tbody>
</table>

- **AI startups of all customer service interactions will be powered by AI bots**
- **$47B spend on AI technologies**
- **20% of companies will dedicate workers to monitor and guide neural networks**
Timeline of AI

1950
Alan Turing proposes the 'Turing Test'

1956
Dartmouth Conference
The modern definitions of AI were defined by Marvin Minsky

1961
First industrial robot (UNIMATE) was introduced at GM

1964
ELIZA, the first chatbot was developed by Weizenbaum at the MIT

1964
AI Winter
False expectations, and limitations in technology left AI out of focus

1997
IBM Deep Blue defeats chess champion Gary Kasparov

2011
IBM Watson beats champions of Jeopardy

2011
The arrival of SIRI

2012
Breakthrough ALEXNET Using NVIDIA GPUs

2014
EUGENE Goostman, a chatbot passes the Turing test. Arrival of Alexa

2015
Google releases Tensorflow

2017
IBM DLL record benchmark with IBM POWER 822LC
Examples and adoptions of AI systems

Automotive, Transportation and Logistics
- Autonomous driving
- Pedestrian detection
- Accident avoidance
- Predictive Maintenance
- Digital twin
- Logistics optimization

Security, Public Safety and Traffic control
- Video Surveillance
- Image analysis
- Facial recognition
- Predictive crime
- Traffic prediction
- Cyber Security

Consumer, Web, Mobile & Retail
- Image tagging
- Speech recognition
- Natural language
- Sentiment analysis
- Recommendation
- Social analysis & trends

Broadcast, Media and Entertainment
- Captioning
- Search
- Recommendations
- Real time translation
- Consumer behaviour

Medicine and Biology
- Drug discovery
- Diagnostic assistance
- Cancer cell detection
- Brain research
- Genome research
- Field studies

Banking, Finance & Insurance
- Trend prediction
- Document analytics
- Recommendation
- Service & Chatbots
- Trading forecast
- Risk management
Challenges of AI

Accuracy
- Data Volume
- Storage Capacity
- Neuronal Network Size

Time
- Compute Power
- Network
- as a Service

Data preparation
- Automation
Sic Transit Gloria Mundi

Google Brain 2012

- 16,000 Servers
- ~ 8 mW/h
- ~ 50 TFLOPS

2015

- 3 NVIDIA PASCAL GPUs
- ~ 0.9 kW/h
- ~ 62 TFLOPS

2017

- 1 NVIDIA Volta GPU
  - ~ 0.3 kW/h
  - ~ 120 TFLOPS
IBM Platform for Deep Learning / Artificial Intelligence

Detect and Collect
- Image & Video
- Text
- Voice & Sound
- Sensor
- CoInt, ELInt, SigInt

Store/Analyze
- Compress/Map Reduce
- Tag/Aggregate
- Knowledge Base

Learn
- Distributed Deep Learning
- Comparison and interpretation
- Combine
- Conclude/Reason

Complementing IBM AI Vision for automation and scaleout DDL

IBM Storage for Analytics & Deep Learning

Analytic Frameworks and solutions:
- Hadoop (Apache)
- Spark (Apache)
- Hortonworks

Filesystems
- IBM Spectrum Scale
- BeeGFS
- CEPH/XFS

IBM Systems and PowerAI Framework

Deep Learning Frameworks
- Caffe
- Chainer
- Torch
- Theano

Supporting Libraries
- OpenBLAS
- NVIDIA DIGITS
- nccl
- Bazel

IBM POWER 822LC
Breakthrough performance for DL/AI and HPC with native NVLINK

Complementing Cloud Services

Platforms
- FPGA
- Applications
- Appliances

Applied Knowledge

Platforms
- FPGA
- Applications
- Appliances
IBM Power Systems LC Line for AI, HPC and BigData
OpenPOWER servers for cloud and cluster deployments that are different by design

S822LC For Big Data
- Ideal for storage-centric and high data through-put workloads
- Brings 2 POWER8 sockets for Big Data workloads
- Big data acceleration with work CAPI and GPUs

S822LC For High Performance Computing
- Incorporates the new POWER8 processor with NVIDIA NVLink
- Delivers 2.8X the bandwidth to GPUs accelerators
- Up to 4 integrated NVIDIA “Pascal” GPUs

S822LC
- 2X memory bandwidth of Intel x86 systems
- Memory Intensive workloads

S821LC
- 2 POWER8 sockets in a 1U form factor
- Ideal for environments requiring dense computing

High Performance Computing
IBM Systems and PowerAI Framework

<table>
<thead>
<tr>
<th>Deep Learning Frameworks:</th>
<th>Caffe</th>
<th>Chainer</th>
<th>torch</th>
<th>theano</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supporting Libraries</td>
<td>OpenBLAS</td>
<td>NVIDIA DIGITS</td>
<td>nccl</td>
<td>Bazel</td>
</tr>
</tbody>
</table>

IBM POWER 822LC
Breakthrough performance for DL/Al and HPC with native NVLINK
IBM Storage for Analytics and Deep Learning

**Analytic Frameworks and solutions:**
- Hadoop
- Apache Spark
- Hortonworks

**Filesystems**
- IBM Spectrum Scale
- BeeGFS
- CEPH/XFS

**IBM Elastic Storage Server (ESS)**
- Extreme Scalability
- Breakthrough performance
- Integrated solution
- IB and Etn Support

- IBM Power System 822
- Scalable technology
- Open Power design
- Linux only
- Flash, SAS SSD
- IB and Etn Support

- IBM Power System CS822
- IBM-NUTANIX appliance
- Hyperconverged Cloud platform
- Flash only (15TB flash/system!)
- NFS
- Etn Support

- IBM-NUTANIX appliance
- OpenPower and Open SDN
- Linux only
- Flash, SAS SSD
- IB and Etn Support
Power AI takes advantage of NVLink between the POWER8 CPU and the P100 GPUs to increase system bandwidth, reduce runtime.

- NVLink only between GPUs
- Long lasting ramp-up times due to PCIe Bottleneck
- Reduced efficiency

- NV Link between CPUs and GPUs enables fast memory access to large data sets in system memory
- Two NVLink connections between each GPU and CPU-GPU leads to faster data exchange
- Distributed Deep Learning (DDL) Record Benchmark
- 3x time saving for learning/training runs in comparison to x86
- Add. CAPI feature for fast IO to storage and network
- Proven scalability up to 256 P100 GPUs in a cluster
Optimizing the development of AI with IBM AI Vision

Typical Challenges in AI projects
- Time consuming, expensive and questionable outcome
- No experience on DNN design and development
- No experience on computer vision
- No experience on how to build a platform to support enterprise scale deep learning, including data preparation, training, and inference

Automation done by IBM AI Vision
- **AI Vision** automates the deep learning development cycles for developers.
- Deep knowledges of ML/DL and computer vision have been embedded into **AI Vision**.
- Reduces time, cost and complexity for AI integration
PowerAI Inference Engine (AccDNN): Automatically generate deep learning accelerator

Automatically enable deep learning from cloud to edge – Enhance productivity

Trained Caffe CNN model in data center

FPGA Accelerator bit-file for edge

Net Model File

Verilog File

FPGA Bit File

FPGA Execution

FPGA chip range from $20 to $1K

Name: "dummy-net"
layers { name: "data" ...}layers { name: "conv" ...}layers { name: "pool" ...}... more layers ...layers { name: "loss" ...}

--- input module ---
conv conv_instance(...)pool pool_instance(...)... more layersloss loss_instance(...)-- output module ---

PowerAI Inference Engine (AccDNN): Automatically enable deep learning from cloud to edge – Enhance productivity

Net.bit
Examples
Mission:
Creating next generations of thinking and self-learning systems based on a deep understanding of cognitive computing and machine learning.

Solutions:
- Traffic Surveillance
- Logistic and Postal Automation
- Document Analysis
- Speech
- Cloud Services
- Mobile Computing
Augmented Working Memory
Neural Turing Machine
Differentiable Neural Computer

Internal Meaning
Representation
Embeddings/Perception Matrix

Convolutional Layer

Recurrent Convolutional Layer
GRU, MDLSTM

Input Sequence

Convolutional Layer

Expectation
1N73LL1G3NC3
15 7h3
4B1L17V
70 4D4P7 70
CH4NG3.

Output Sequence
Beam Search

SEQUENCE-TO-SEQUENCE
END-TO-END TRAINABLE
IBM POWER 822LC 4 x P100 GPU
150 TFLOPs
benchmarks with
- speech
- handwriting
- visual object recognition
600 times faster than CPU
Use cases of PlanetBrain

- Traffic
- Logistic
- Document Analysis
Traffic

Planet software based on PlanetBrain is:
- finding and tracking vehicles
- reading number plate
- finding driver face
- drop all if beautiful girl is driving
Traffic

- success rate: 97%
- processing in real-time in CPU
- approx. 400 systems in Germany, Austria, Switzerland
129 km/h

<table>
<thead>
<tr>
<th>Nr.</th>
<th>Zeit</th>
<th>Geschwindigkeit</th>
<th>Abstand</th>
<th>x/10</th>
<th>Bild</th>
</tr>
</thead>
<tbody>
<tr>
<td>0001</td>
<td>01:37:08</td>
<td>141 km/h</td>
<td></td>
<td></td>
<td>0577</td>
</tr>
<tr>
<td>0002</td>
<td>01:37:17</td>
<td>108 km/h</td>
<td></td>
<td></td>
<td>0801</td>
</tr>
<tr>
<td>0003</td>
<td>01:37:20</td>
<td>151 km/h</td>
<td></td>
<td></td>
<td>0855</td>
</tr>
<tr>
<td>0004</td>
<td>01:37:21</td>
<td>151 km/h</td>
<td></td>
<td></td>
<td>0887</td>
</tr>
<tr>
<td>0005</td>
<td>01:37:22</td>
<td>129 km/h</td>
<td></td>
<td></td>
<td>0922</td>
</tr>
</tbody>
</table>

https://www.facebook.com/pg/PlanetAIgmbh/videos/
Planet software based on PlanetBrain is:

- finding Regions of Interest (ROI)
- reading address fields
- distinguishing between receiver and sender
Logistic

success rate: 85% - 97%
processing time: 0.2 - 5 sec on CPU

USA: several hundred systems at FedEx and USPS
Europe: > 10 large mail distributors
Logistic

https://www.facebook.com/pg/PlanetAlGmbH/videos/
Document Analysis

Automatic inbox processing:
- converting paper documents into classified PDF (as email attachment)
- processing 50,000 documents per hour on a single PowerAI machine

Solutions:
- Insurance
- Healthcare
- Finance
- Government
Sehr geehrtes AOK-Team,

dient bitte ich

859, um eine Rückerstattung von der freiwilligen Krankenversicherung von Mai 2015-Nov. 2015 auf den folgenden Kontonr: IBAN: DE 58 100 70848 

BIC: DEUTDEDB110

Mit freundlichen Grüßen

Berlin, den 18.11.15
Document Analysis

reading handwritten and machine printed documents

- processing time: 10 sec / page / CPU
- READ: the largest EU project (H2020) European Cultural Heritage
  11 billion pages 1500 - 1800
ArgusSearch in handwriting

https://www.facebook.com/pg/PlanetAlGmbH/videos/
ArgusSearch in speech

https://www.facebook.com/pg/PlanetAIgmbh/videos/
AlaaS
About INS group

• Founded: 1992
• Managed IT services
• IT-outsourcing
• Data center operation
• Cloud services
• Hosting
• Network & security
• Software as a Service
• Procurement

Founded: 2005
• IT service desk
• User help desk
• Technical services
• Service hotlines

Technology consultancy
Process consultancy
IT projects
Business Process Management

TIER 3+ Data Centers in Hanover, Frankfurt/Main, Lucerne (CH)
Challenges

- You wish to try out the technology within a Proof of Concept (POC)?
- You only require resources temporarily?
- You need scalable and flexible resources?
- You don’t want to worry about security and compliance issues?
- You don’t want outlays in regards to backup or operation?
- …

Execute your Cognitive Computing applications on servers which were explicitly developed for such a task. We can assist you with our resources.

Competent, flexible and straight-forward.
Service model – Platform as a Service

Docker application containers
Docker container management tool as a tenant
Data will be provided physical or from within the cloud
Connection via VPN, SFTP or HTTPS
Appropriate NFS storage
Additional temporary storage can be added at any time
Availability and backup SLA
Configuration IBM Power 822LC HPC

- **IBM Power 822LC HPC**
  - **IB EDR Adapter**
  - **2 * 100 Gbit**
  - **On Board 4 * 10 Gbit Etn**
  - **PEX/CAPI**
  - **SSD or SAS**

### CPU 1: POWER 8+ 8 or 10Core
- **32 GB**
- **NVIDIA TESLA® 100 GPU**
- **NVLINK 40GB + 40GB bidirectional**
- **POWER8 SMP-A 3 x 12,8GB/s**
- **4 Lanes / CPU (115GB/s per CPU)**

### CPU 2: POWER 8+ 8 or 10Core
- **32 GB**
- **NVIDIA TESLA® 100 GPU**
- **NVLINK 40GB + 40GB bidirectional**

### Memory
- **16GB**
- **4 x NVIDIA TESLA® 100 GPU**
- **NVMe 1.6TB**

---

**Insider Knowledge**

- **Technology - Communication - Consulting**

**IBM Power Systems**

**IBM Cognitive Computing PaaS**
Setup / System configuration

1. OPEX based operating models:
   a. Pay per use based on INS platform services.
   b. Individual Cloud based Datacenter configurations on long term contracts.
   c. On Premise installations of HPC cluster systems combined with Managed Services by INS.

2. CAPEX and OPEX combined models:
   a. On Premise installations of HPC cluster systems combined with Managed Services by INS.
   b. On Premise delivery in individual configurations based on customer requirements

Typical system configurations are:

- **Management System**: usually VM
- **Monitoring Satellite**: System Monitoring (usually VM)
- **IBM Cloud Private System**: usually VM
- **Storage Connector System**: based on NFS à Based on ordered storage type (physical server / system or VM or combined system)
- **IBM Power S822LC system**: Compute nodes 1 … n
- **Networking**: 10Gbe up to InfiniBand 100Gbe connections possible
  Connections based on requirements by systems.
  Uplink 1000BaseT up to 100Gbe
Connecting data islands for a hyperconnected and cognitive universe

- Security, defence, protection of cyber crime
- Health & research
- Weather, climate research & Agriculture
- car2X, autonomous vehicles and intelligent traffic systems
- Wearables & mobility, Infotainment, industrial & military health and fitness
- Industry 4.0
- Banking, finance & insurance
- Energy, utilities and Smart cities
- Connected Home
- Retail and Marketing
Legal Notices

Copyright © 2016 by International Business Machines Corporation. All rights reserved.

No part of this document may be reproduced or transmitted in any form without written permission from IBM Corporation.

Product data has been reviewed for accuracy as of the date of initial publication. Product data is subject to change without notice. This document could include technical inaccuracies or typographical errors. IBM may make improvements and/or changes in the product(s) and/or program(s) described herein at any time without notice. Any statements regarding IBM’s future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only. References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business. Any reference to an IBM Program Product in this document is not intended to state or imply that only that program product may be used. Any functionally equivalent program, that does not infringe IBM's intellectually property rights, may be used instead.

THE INFORMATION PROVIDED IN THIS DOCUMENT IS DISTRIBUTED “AS IS” WITHOUT ANY WARRANTY, EITHER OR IMPLIED. IBM LY DISCLAIMS ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NONINFRINGEMENT. IBM shall have no responsibility to update this information. IBM products are warranted, if at all, according to the terms and conditions of the agreements (e.g., IBM Customer Agreement, Statement of Limited Warranty, International Program License Agreement, etc.) under which they are provided. Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products in connection with this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. IBM makes no representations or warranties, ed or implied, regarding non-IBM products and services.

The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents or copyrights. Inquiries regarding patent or copyright licenses should be made, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-785
U.S.A.
Legal Notices

IBM, the IBM logo, ibm.com, IBM System Storage, IBM Spectrum Storage, IBM Spectrum Control, IBM Spectrum Protect, IBM Spectrum Archive, IBM Spectrum Virtualize, IBM Spectrum Scale, IBM Spectrum Accelerate, Softlayer, and XIV are trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. A current list of IBM trademarks is available on the Web at “Copyright and trademark information” at http://www.ibm.com/legal/copytrade.shtml

The following are trademarks or registered trademarks of other companies.
Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.
IT Infrastructure Library is a Registered Trade Mark of AXELOS Limited.
Linear Tape-Open, LTO, the LTO Logo, Ultrium, and the Ultrium logo are trademarks of HP, IBM Corp. and Quantum in the U.S. and other countries.
Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.
Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.
Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.
Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.
Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license therefrom.
ITIL is a Registered Trade Mark of AXELOS Limited.
UNIX is a registered trademark of The Open Group in the United States and other countries.
* All other products may be trademarks or registered trademarks of their respective companies.

Notes:
Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user’s job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.

All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.
This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.
All statements regarding IBM’s future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.
Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the performance, compatibility, or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.

This presentation and the claims outlined in it were reviewed for compliance with US law. Adaptations of these claims for use in other geographies must be reviewed by the local country counsel for compliance with local laws.