Meeting of the LF AI & Data Technical Advisory Council (TAC)

September 23, 2021
Recording of Calls

Reminder:

TAC calls are recorded and available for viewing on the TAC Wiki
Reminder: LF AI & Data Useful Links

› Web site: lfaidata.foundation
› Wiki: wiki.lfaidata.foundation
› GitHub: github.com/lfaidata
› Landscape: https://landscape.lfaidata.foundation or https://l.lfaidata.foundation
› Mail Lists: https://lists.lfaidata.foundation
› Slack: https://slack.lfaidata.foundation
› Youtube: https://www.youtube.com/channel/UCfasaeqXJBCAJMNO9HcHfbA
› LF AI Logos: https://github.com/lfaidata/artwork/tree/master/lfaidata
› LF AI Presentation Template: https://drive.google.com/file/d/1eiDNJvXCqSZHT4ZkczASl2GTBRZk2/view?usp=sharing
› Events Page on LF AI Website: https://lfaidata.foundation/events/
› Events Calendar on LF AI Wiki (subscribe available): https://wiki.lfaidata.foundation/pages/viewpage.action?pageId=12091544
› Event Wiki Pages: https://wiki.lfaidata.foundation/display/DL/LF+AI+Data+Foundation+Events
Agenda

› Roll Call
› Approval of Minutes from previous meetings
› Integration Update: Triton, Milvus, and Feast
› NNStreamer Project Update
› LF AI General Updates
› Open Discussion
<table>
<thead>
<tr>
<th>Board Member</th>
<th>Contact Person</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT&amp;T</td>
<td>Anwar Atfab*</td>
<td><a href="mailto:anwar@research.att.com">anwar@research.att.com</a></td>
</tr>
<tr>
<td>Baidu</td>
<td>Ti Zhou</td>
<td><a href="mailto:zhouti@baidu.com">zhouti@baidu.com</a></td>
</tr>
<tr>
<td>Ericsson</td>
<td>Rani Yadav-Ranjan*</td>
<td><a href="mailto:rani.yadav-ranjan@ericsson.com">rani.yadav-ranjan@ericsson.com</a></td>
</tr>
<tr>
<td>Huawei</td>
<td>Huang Zhipeng</td>
<td><a href="mailto:huangzhipeng@huawei.com">huangzhipeng@huawei.com</a></td>
</tr>
<tr>
<td>IBM</td>
<td>Susan Malaika</td>
<td><a href="mailto:malaika@us.ibm.com">malaika@us.ibm.com</a></td>
</tr>
<tr>
<td>Nokia</td>
<td>Jonne Soininen</td>
<td><a href="mailto:jonne.soininen@nokia.com">jonne.soininen@nokia.com</a></td>
</tr>
<tr>
<td>OPPO</td>
<td>Jimin Jia*</td>
<td><a href="mailto:jiajimin@oppo.com">jiajimin@oppo.com</a></td>
</tr>
<tr>
<td>SAS</td>
<td>Nancy Rausch</td>
<td><a href="mailto:nancy.rausch@sas.com">nancy.rausch@sas.com</a></td>
</tr>
<tr>
<td>Tech Mahindra</td>
<td>Amit Kumar</td>
<td><a href="mailto:Kumar_Amit@techmahindra.com">Kumar_Amit@techmahindra.com</a></td>
</tr>
<tr>
<td>Tencent</td>
<td>Bruce Tao</td>
<td><a href="mailto:brucetao@tencent.com">brucetao@tencent.com</a></td>
</tr>
<tr>
<td>Zilliz</td>
<td>Jun Gu</td>
<td><a href="mailto:jun.gu@zilliz.com">jun.gu@zilliz.com</a></td>
</tr>
<tr>
<td>ZTE</td>
<td>Wei Meng</td>
<td><a href="mailto:meng.wei2@zte.com.cn">meng.wei2@zte.com.cn</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Graduate Project</th>
<th>Contact Person</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acumos</td>
<td>Nat Subramanian</td>
<td><a href="mailto:natarajan.subramanian@techmahindra.com">natarajan.subramanian@techmahindra.com</a></td>
</tr>
<tr>
<td>Angel</td>
<td>Bruce Tao</td>
<td><a href="mailto:brucetao@tencent.com">brucetao@tencent.com</a></td>
</tr>
<tr>
<td>Egeria</td>
<td>Mandy Chessell</td>
<td><a href="mailto:mandy_chessell@uk.ibm.com">mandy_chessell@uk.ibm.com</a></td>
</tr>
<tr>
<td>Horovod</td>
<td>Travis Addair*</td>
<td><a href="mailto:taddair@uber.com">taddair@uber.com</a></td>
</tr>
<tr>
<td>Milvus</td>
<td>Xiaofan Luan</td>
<td><a href="mailto:xiaofan.luan@zilliz.com">xiaofan.luan@zilliz.com</a></td>
</tr>
<tr>
<td>ONNX</td>
<td>Jim Spohrer (Chair of TAC)</td>
<td><a href="mailto:spohrer@us.ibm.com">spohrer@us.ibm.com</a></td>
</tr>
<tr>
<td>Pyro</td>
<td>Fritz Obermeyer*</td>
<td><a href="mailto:fritz.obermeyer@gmail.com">fritz.obermeyer@gmail.com</a></td>
</tr>
</tbody>
</table>
Approval of August 26th, 2021 Minutes

Draft minutes from the August 26th TAC call were previously distributed to the TAC members via the mailing list

Proposed Resolution:
› That the minutes of the August 26th meeting of the Technical Advisory Council of the LF AI & Data Foundation are hereby approved.
Approval of September 9th, 2021 Minutes

Draft minutes from the September 9th TAC call were previously distributed to the TAC members via the mailing list.

Proposed Resolution:
› That the minutes of the September 9th meeting of the Technical Advisory Council of the LF AI & Data Foundation are hereby approved.
Update on Integration: Triton, Milvus, Feast

September 23, 2021

Jun Gu
NVIDIA Merlin addresses Recommender System challenges

**ETL**
- GPUs are slow and complex
- GPU-accelerated and easy-to-use ETL pipelines prepare datasets in minutes

**Data Loading**
- Using common item-by-item loading can be slow
- Asynchronous and GPU-accelerated dataloader for PyTorch and TensorFlow/Keras

**Training**
- Embedding tables of large deep learning recommender systems can exceed memory
- Easy data and model parallel training allow to scale TB size embeddings

**Inference**
- High throughput to rank more items is difficult while maintaining low latency
- High throughput, low-latency production deployment

NVIDIA Merlin is an open-source library to deploy recommender systems end-to-end
<table>
<thead>
<tr>
<th>Company</th>
<th>Use Case</th>
<th>Workflows</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walmart</td>
<td>Accelerating real-time dynamic pricing for in-store grocery items to avoid waste</td>
<td>Merlin NVTabular, Dask and TF to scale data preprocessing Multi-GPU train a wide and deep model</td>
<td>10X ETL Operations Speedup With Merlin NVTabular</td>
</tr>
<tr>
<td>Tencent</td>
<td>Increased accuracy in advertising recommendations for 1B MAU</td>
<td>Integration of HugeCTR framework into advertising recommendation training framework</td>
<td>Training time reduced from 20 hrs to 3 7x speed-up in comparison to the previous TF solution using the same GPU infrastructure</td>
</tr>
<tr>
<td>Snapchat</td>
<td>Recommenders curate real-time content feeds, encourage inspirational sharing</td>
<td>Using NVIDIA T4 GPUs, Triton and TensorRT</td>
<td>DL inference cost efficiency by 50% and decrease serving latency by over 60%</td>
</tr>
<tr>
<td>Pinterest</td>
<td>Provide a delightful customer experience through highly relevant search results and recommendations</td>
<td>SotA CV uses Triton on V100 to process 300B pins(images) every night to extract embedded features</td>
<td>3x higher throughput with Triton dynamic batching</td>
</tr>
<tr>
<td>Uber</td>
<td>Personalization and discovery features for three sided marketplace</td>
<td>RAPIDS.AI and Merlin NVTabular for TF Dataloader</td>
<td>Training time reduced from 1 hr (CPU) to 5 min (GPU) Reduced cost by 95% on NVIDIA A100 GPUs</td>
</tr>
<tr>
<td>Spotify</td>
<td>Improve playlist recommendations</td>
<td>TensorFlow Model Analysis (TFMA) - Improved with RAPIDS and NVTabular</td>
<td>Model validation time reduced from 3.5 hours (CPU) To 10 secs (GPU)</td>
</tr>
</tbody>
</table>
Four Stage Rec Sys Inference POC w/ Triton, Milvus and Feast

● Goal
  ○ The prototype leverages Triton model ensemble to perform four-stage inference for recommender model integrating Feast for feature retrieval and Milvus for candidate generation

● Functionality
  ○ The prototype takes a user id from the incoming request at inference and return an ordered list of recommendations

● Stages
  ○ Retrieval - nearest neighbor search w/ Milvus (Faiss-GPU)
  ○ Filtering - remove items from current session
  ○ Scoring - predict likelihood of item interaction
  ○ Ordering - softmax exploration

● External components
  ○ Milvus - open-source vector database for similarity search. Uses FAISS as as one of ANN algorithms (FAISS, Annoy, HNSWlib)
  ○ Feast - open-source feature store for model training and online inference
Four-stage Recommender Systems

1. **Online**
   - **Retrieval**
     - Embed input item or query
     - Embedding model
   - **Scoring**
     - Add features to candidates
     - Feature Store
   - **Ordering**
     - Define Ordering Policy

2. **Offline**
   - **Retrieval**
     - Embed items from catalog
     - Embedding model
   - **Filtering**
     - Build Bloom Filters
     - Interaction data
   - **Scoring**
     - Train ranking model
     - Training data
   - **Ordering**
     - Build Bloom Filters
     - Interaction data

3. **Offline-Online**
   - **Retrieval**
     - Build Approx. NN index
   - **Filtering**
     - Filter invalid candidates
     - Bloom Filter
   - **Scoring**
     - Score top k candidates
     - Ranking model
   - **Ordering**
     - Ordering Policy

- **Training data**
- **Item data**
- **Interaction data**
- **Item & user data**
- **Ordering Business logic**

- **Build Feature Store**
- **Build Bloom Filters**
- **Filter invalid candidates**
- **Score top k candidates**
- **Ordering Policy**
- **Define Ordering Policy**

- **Train embedding**
- **Embed items from catalog**
- **Build Approx. NN index**
- **Add features to candidates**
- **Score top k candidates**

- **Results**
- **Items**
- **Candidates**
- **Candidates**
POC diagram with Triton, Milvus and Feast
POC diagram with Triton, Milvus and Feast
Annual Review for NNStreamer

MyungJoo Ham (myungjoo.ham@gmail.com)
9/23/2021
NNStreamer

**Brief Description:**
NNStreamer is a set of Gstreamer plugins that allow Gstreamer developers to adopt neural network models easily and efficiently and neural network developers to manage neural network pipelines and their filters easily and efficiently.

**Contributed by:**
Samsung in March 2020 as an Incubation Project

**Key Links:**
- Github: [https://github.com/nnstreamer/nnstreamer](https://github.com/nnstreamer/nnstreamer)
- Mailing lists:
  - nnstreamer-announce
  - nnstreamer-technical-discuss
  - nnstreamer-tsc
Incubation/Graduation Project review criteria

To be accepted into the Graduation stage, a project must meet the Incubation stage requirements plus:

- Need contributors from more organizations.
- Have reached 380 stars on GitHub.
- *A substantial ongoing flow of commits and merged contributions for the past 12 months*.
- Receive the affirmative vote of two-thirds of the TAC and the affirmative vote of the Governing Board.
- Have completed at least one collaboration with another LF AI & Data hosted project
- Have a technical lead appointed for representation of the project on the LF AI & Data Technical Advisory Council.
### Organizations contributing

<table>
<thead>
<tr>
<th># Contributors</th>
<th># Commits</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt; 20</td>
<td>&gt; 2000</td>
</tr>
<tr>
<td>1</td>
<td>19</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>12 (A. Arthurs, J. Lee, C. Hall, W. Lee, D. Lee, C. Jeon, B. Kim, D. Kim, N. Jang, H. Park)</td>
<td>&gt; 30</td>
</tr>
</tbody>
</table>

Counting nnstreamer.git only (excluding other git repos) 2021-09-14
Contributions

810.05K
Lines Of Code Changed

2.12K
Commits

48
Contributors

1
No Of Sub Projects

10
Repositories

Top 10 Contributors By Commits

<table>
<thead>
<tr>
<th>NAME</th>
<th>LINES OF CODE</th>
<th>COMMITS</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jihoon Lee</td>
<td>114.03K</td>
<td>406</td>
<td>14.08%</td>
</tr>
<tr>
<td>Parichay Kapoor</td>
<td>113.81K</td>
<td>390</td>
<td>14.06%</td>
</tr>
<tr>
<td>Jaeyun Jung</td>
<td>164.96K</td>
<td>291</td>
<td>20.35%</td>
</tr>
<tr>
<td>gichan-jang</td>
<td>142.33K</td>
<td>155</td>
<td>17.57%</td>
</tr>
<tr>
<td>HyoungJoo Ahn</td>
<td>66.79K</td>
<td>147</td>
<td>8.25%</td>
</tr>
<tr>
<td>wooksong</td>
<td>25.14K</td>
<td>140</td>
<td>3.10%</td>
</tr>
<tr>
<td>Jijoong Moon</td>
<td>33.20K</td>
<td>117</td>
<td>4.10%</td>
</tr>
<tr>
<td>MyungJoo Ham</td>
<td>40.26K</td>
<td>99</td>
<td>4.97%</td>
</tr>
<tr>
<td>hsiang-hoo</td>
<td>11.75K</td>
<td>21</td>
<td>1.45%</td>
</tr>
</tbody>
</table>

Top 10 Organizations By Commits

2.12K
Commits

2021-09-16, last 1 year
Key Achievements in the past year

Commercialization / Deployment
- Samsung
  - 2020: Galaxy Watch 3
  - 2021: Bixby Service in Galaxy Series
  - 2022 (WIP): TV, AR apps, Robotics, Home appliances
- Companies contacted for their products: NXP, Dell (Pravega), Collabora, Huawei, ODKMedia, FAInders,
- Tizen (5.5, 6.0, 6.5), Android JCenter, Yocto (meta-neural-network), Ubuntu (PPA), MacOS (Homebrew)

Features & Adaptation
- More stream data types: flexible-tensor, sparse-tensor, flexbuf
- More stream path manipulators: IF-branch, Join, Rate, Crop,
- Adaptation: TF-lite delegation, Tensor-RT (Nvidia), TVM, Lua scripts
- Gst-to-Pbtxt(mediapipe) converter. (for WYSIWYG pipeline editor in the future)

New Concept: Edge-AI (“Among-Device AI”)
- MQTT Pub/Sub, Query Server/Client (Prototype)

New Subproject: NNTrainer (On-Device Neural Network Training)

Areas the project could use help on

- Linux distro deployment
  - Deploy to Debian, OpenSuse, Fedora, ...

- Adding more automated test suites
  - E.g., Coverity, ARM-cloud for testing ARM binaries, ...

- Connection with Matter
  - 2.0+ (2022) features are related w/ “Home IoT Connectivity”.
  - We’d like to contribute “Inter-device AI stream protocol”
    - Possibly, w/ plugins or libraries for MediaPipe, DeepStream
TAC Open Discussion
Upcoming TAC Meetings
Upcoming TAC Meetings (Tentative)

- Oct 7, 2021: Ludwig Annual project review, Amundsen Annual project review; OC update
- Oct 21, 2021: AI Fairness 360, AI Explainability 360, Adversarial Robustness Toolbox Annual project reviews

Please send agenda topic requests to tac-general@lists.lfai-data.foundation
LF AI & Data - Ongoing Annual Project Reviews
<table>
<thead>
<tr>
<th>Date</th>
<th>Project</th>
<th>Presenter</th>
</tr>
</thead>
<tbody>
<tr>
<td>April 4, 2021</td>
<td>Sgeria</td>
<td>Molly Chevalier (slack) - TAC/record</td>
</tr>
<tr>
<td>April 6, 2021</td>
<td>OpenDial</td>
<td>Andre de Waal (slack) - TAC/record</td>
</tr>
<tr>
<td>May 20, 2021</td>
<td>DNNIE</td>
<td>Jim Spohrer (slack) - TAC/record</td>
</tr>
<tr>
<td>July 15, 2021</td>
<td>ESL</td>
<td>Ti Zhou (slack) - deck</td>
</tr>
<tr>
<td>July 29, 2021</td>
<td>Angel</td>
<td>Bruce Tan (slack) - confirmed deck</td>
</tr>
<tr>
<td>July 29, 2021</td>
<td>Addik</td>
<td>Meng Wei (slack) - confirmed deck</td>
</tr>
<tr>
<td>Aug 12, 2021</td>
<td>(potentially Aug 13)</td>
<td>Sparky</td>
</tr>
<tr>
<td>Aug 13, 2021</td>
<td>Miles</td>
<td>Jun Gu (slack) - confirmed</td>
</tr>
<tr>
<td>Aug 24, 2021</td>
<td>Kendro new project into incubation</td>
<td>Yekunde Dada - <a href="mailto:yekunde_dada@newsify.com">yekunde_dada@newsify.com</a></td>
</tr>
<tr>
<td>Sept 6, 2021</td>
<td>Marquee</td>
<td>Julian Le Dain (slack) - confirmed</td>
</tr>
<tr>
<td>Sept 6, 2021</td>
<td>Acrinos</td>
<td>Amit Kumar (slack) - tentative</td>
</tr>
<tr>
<td>Sept 21, 2021</td>
<td>NKEStreamer</td>
<td>Myungwoo Kim (slack) - confirmed</td>
</tr>
<tr>
<td>Sept 21, 2021</td>
<td>Forediow</td>
<td>Ahmed Abbadi (slack) - confirmed</td>
</tr>
<tr>
<td>Oct 7, 2021</td>
<td>Ludwig</td>
<td>Riess Molinov (slack) - confirmed</td>
</tr>
<tr>
<td>Oct 7, 2021</td>
<td>Amundsen</td>
<td>Mark Grover (slack) - confirmed</td>
</tr>
<tr>
<td>Oct 21, 2021</td>
<td>AI Fairness 360</td>
<td>Anshu Singh (to be asked)</td>
</tr>
<tr>
<td>Oct 21, 2021</td>
<td>AI Explainability 360</td>
<td>Anshu Singh (to be asked)</td>
</tr>
<tr>
<td>Oct 21, 2021</td>
<td>Adversarial Robustness Toolbox</td>
<td>Anshu Singh (to be asked)</td>
</tr>
<tr>
<td>Nov 4, 2021</td>
<td>Harrood</td>
<td>Travis Adair (to be asked)</td>
</tr>
<tr>
<td>Nov 4, 2021</td>
<td>REACT</td>
<td>William Freeman (to be asked)</td>
</tr>
<tr>
<td>Nov 10, 2021</td>
<td>SQUIS</td>
<td>Anthony Lage (to be asked)</td>
</tr>
<tr>
<td>Nov 10, 2021</td>
<td>Delta</td>
<td>Kun Hee (to be asked)</td>
</tr>
<tr>
<td>Dec 2, 2021</td>
<td>DataPractice.org</td>
<td>Patrice McInerney (to be asked)</td>
</tr>
<tr>
<td>Dec 2, 2021</td>
<td>LenaGraph</td>
<td>Jason Plural (to be asked)</td>
</tr>
<tr>
<td>Dec 16, 2021</td>
<td>Pyro</td>
<td>Fritz Obermayer (to be asked)</td>
</tr>
<tr>
<td>Jan 6, 2022</td>
<td>Datasheets</td>
<td>Yannick Goudouze (to be asked)</td>
</tr>
<tr>
<td>Jan 6, 2022</td>
<td>Ryte</td>
<td>Kefan Omar (to be asked)</td>
</tr>
<tr>
<td>Jan 20, 2022</td>
<td>ROSAIS</td>
<td>Ludan Sleedik (to be asked)</td>
</tr>
<tr>
<td>Jan 20, 2022</td>
<td>SubstraFramework</td>
<td>Camille Marin (to be asked)</td>
</tr>
<tr>
<td></td>
<td>Machine Learning Exchange</td>
<td>Anshu Singh (to be asked)</td>
</tr>
<tr>
<td></td>
<td>VulcanoCompute</td>
<td>Alejandro Saondo (to be asked)</td>
</tr>
<tr>
<td></td>
<td>OpenLineage</td>
<td>Julian Le Dain (to be asked)</td>
</tr>
<tr>
<td></td>
<td>MARS</td>
<td>Chris Qin (to be asked)</td>
</tr>
</tbody>
</table>

**Schedule:** [https://wiki.lfaidata.foundation/pages/editpage.action?pageId=43286684](https://wiki.lfaidata.foundation/pages/editpage.action?pageId=43286684)
TAC Meeting Details

› To subscribe to the TAC Group Calendar, visit the wiki:
  https://wiki.lfaidata.foundation/x/cQB2

› Join from PC, Mac, Linux, iOS or Android: https://zoom.us/j/430697670

› Or iPhone one-tap:
  › US: +16465588656,,430697670# or +16699006833,,430697670#

› Or Telephone:
  › Dial(for higher quality, dial a number based on your current location):
    › US: +1 646 558 8656 or +1 669 900 6833 or +1 855 880 1246 (Toll Free) or +1 877 369 0926 (Toll Free)

› Meeting ID: 430 697 670

› International numbers available: https://zoom.us/u/achYtcw7uN
Open Discussion
Legal Notice

- The Linux Foundation, The Linux Foundation logos, and other marks that may be used herein are owned by The Linux Foundation or its affiliated entities, and are subject to The Linux Foundation’s Trademark Usage Policy at https://www.linuxfoundation.org/trademark-usage, as may be modified from time to time.

- Linux is a registered trademark of Linus Torvalds. Please see the Linux Mark Institute’s trademark usage page at https://lmi.linuxfoundation.org for details regarding use of this trademark.

- Some marks that may be used herein are owned by projects operating as separately incorporated entities managed by The Linux Foundation, and have their own trademarks, policies and usage guidelines.

- TWITTER, TWEET, RETWEET and the Twitter logo are trademarks of Twitter, Inc. or its affiliates.

- Facebook and the “f” logo are trademarks of Facebook or its affiliates.

- LinkedIn, the LinkedIn logo, the IN logo and InMail are registered trademarks or trademarks of LinkedIn Corporation and its affiliates in the United States and/or other countries.

- YouTube and the YouTube icon are trademarks of YouTube or its affiliates.

- All other trademarks are the property of their respective owners. Use of such marks herein does not represent affiliation with or authorization, sponsorship or approval by such owners unless otherwise expressly specified.

- The Linux Foundation is subject to other policies, including without limitation its Privacy Policy at https://www.linuxfoundation.org/privacy and its Antitrust Policy at https://www.linuxfoundation.org/antitrust-policy, each as may be modified from time to time. More information about The Linux Foundation’s policies is available at https://www.linuxfoundation.org.

- Please email legal@linuxfoundation.org with any questions about The Linux Foundation’s policies or the notices set forth on this slide.

[Header: LF AI & DATA]