Transcript
NSF- RTML (NSF 19-566) Webinar
April 3, 2019, 3:00pm EST

slide 1
<Rance Cleaveland>
Welcome to today's National Science Foundation webinar on the Real-Time Machine Learning, or RTML, program. My name is Rance Cleaveland, and I am the Division Director for the Computing and Communication Foundations, or CCF, research division within the Computing and Information Science and Engineering Directorate at NSF.

slide 2
<Rance Cleaveland>
I would also like to introduce three other NSF people who will be speaking today as well. These are: Dr. Fil Bartoli, who is the Division Director for the Electrical, Communications and Cyber Systems, or ECCS, research division within the Engineering Directorate at NSF; Dr. Sankar Basu, Program Director in CCF; and Dr. Jenshan Lin, Program Director in ECCS. Drs. Basu and Lin are the Cognizant Program Directors for the RTML program.

The rest of the webinar will proceed as follows. After introductory remarks by me and Dr. Bartoli, Drs. Basu and Lin will provide an overview of the RTML solicitation. They will then entertain your questions until 4 pm Eastern Time, or until all questions are answered, whichever comes first.

slide 3
<Rance Cleaveland>
If you have any questions about the solicitation after this meeting ends, please feel free to e-mail them to Drs. Basu and Lin; their e-mail addresses are on the slide. Also, a video recording of this webinar, as well as the slides used, will be available on the CISE web-site at the indicated address.

slide 4
<Rance Cleaveland>
The RTML program is a cross-cutting program that involves participation of two research divisions – CCF and ECCS – from two different directorates – CISE and Engineering. It also involves a teaming arrangement with the Defense Advanced Research Projects Agency, or DARPA, though a separately managed but contemporaneous DARPA program. I will speak briefly about the importance of the topic from the CCF perspective; then, Dr. Bartoli will discuss ECCS's interests in the program as well.
Let me first talk briefly about CCF. This division has four so-called "clusters," or areas of focus: Algorithmic Foundations, Communication and Information Foundations, Foundations of Emerging Technologies, such as quantum and bio-computing, and Software and Hardware Foundations.

Interest in Machine Learning, or ML, has grown explosively over the past few years, driven by the wide-ranging applications in many different areas of science and engineering but also other

fields. ML often relies on the processing of massive amounts of data, often continuously streaming, for training purposes, and this leads to questions about how to support these applications most efficiently, in terms of time but also energy consumption, from both a hardware and an algorithmic perspective. The CCF division that I lead has algorithmic foundations and hardware foundations as two of its focus areas, and thus is a natural fit for a program, like RTML, focused on developing transformative approaches to algorithmic and hardware co-design for machine-learning applications. Indeed, the depth and breadth of these applications, and the widely varying constraints they impose in terms of time and energy consumption, require the research community to re-think the foundational aspects, both algorithmic and hardware, for implementing ML. I think these topics are very exciting, and I am looking forward to the research results that come out of this program.

Let me now turn the floor over to Dr. Fil Bartoli, Division Director of ECCS.

slide 5
<Fil Bartoli>
Let me also welcome you on behalf of the Electrical, Communications and Cyber Systems Division (ECCS). Our Division is organized into the following three Clusters: Electronics, Photonics and Magnetic Devices (EPMD), Communications, Circuits, and Sensing Systems (CCSS), Energy, Power, Control and Networks (EPCN). Machine learning activities in ECCS are primarily concentrated in the CCSS and EPCN Clusters.

Machine-Learning algorithms and advanced electronic hardware have long been of interest to the ECCS community, involving researchers in areas such as control, signal processing, information theory, and microelectronics. Recent work on ML algorithms that employ "gradient descent" to establish cost function is an extension of prior work in electrical engineering on adaptive filters. ML has achieved some success in recent years due to the increased availability of GPUs and access to large data sets. However, the implementation in dynamic engineering systems that directly address societal challenges will require significant advances in real-time sensing, learning and distributed decision-making over a wide range of applications, such as autonomous vehicles, health technologies, defense systems and manufacturing. ECCS is delighted to work with you in promoting foundational and domain-specific advances in real-time learning to meet these exciting challenges.

Let me now turn the floor over to Dr. Sankar Basu, Program Director in the CISE/CCF Division.

slide 6
<Sankar Basu>

Thank you Dr. Bartoli.

It has been said for several years now that a grand challenge in computing is the "creation of machines that can proactively interpret and learn from data in real time, solve unfamiliar problems using what they have learned, and operate with the energy efficiency of the human brain". The genesis of this statement goes back to a computing grand challenge announced by the White House OSTP back in 2015.

The webpage for this Grand Challenge is now archived, but is available by a google search of the string "A Nanotechnology-Inspired Grand Challenge for Future Computing" (also, the transcript of this webcast will have that URL listed)

URL:
https://obamawhitehouse.archives.gov/blog/2015/10/15/nanotechnology-inspired-grand-challenge-future-computing

While prior NSF solicitations along this line focused on Energy Efficiency, the current solicitation, together with DARPA, focusses on (machine) learning hardware with emphasis on real time aspects.

The program is also inspired by the success of machine learning algorithms in several areas, e.g., in speech recognition/computer vision etc., and at the same time emergence of new challenges with greater emphasis on real time applications in broad areas of autonomous vehicles, and 5G applications.

The purpose of this webcast is to inform our PI community about the NSF-RTML Program solicitation, describe collaboration with the DARPA-RTML broad agency announcement, and also to respond to questions from relevant parties such as potential applicants, i.e., PIs and university grant administrators.

slide 7
<Sankar Basu>

The main goal of the joint program with NSF and DARPA is to explore rapid development of efficient hardware architectures for Machine Learning (abbreviated as ML) from continuous stream of real time data.

One of the DARPA (as well as NSF) interests is also to consider distributed Machine Learning (ML) in a cloud environment.

There are two target application areas of interest to DARPA inspired by autonomous vehicles scenario, which are design of:

1. High bandwidth image processing systems constrained by Size, Weight and Power
2. High bandwidth wireless systems such as 60 GHz for 5G applications.

DARPA may at some point provide training and/or test data for at least one of these two applications.

It must be understood, however, that
aim of the NSF-RTML program is primarily in keeping with NSF mission of basic/fundamental research,

whereas

the goal of the DARPA-RTML program is to create the tools and circuit development infrastructure enabling rapid innovation in resulting Machine Learning hardware – possibly by leveraging the other ongoing DARPA programs on chip-design.

That said, collaboration between the two agencies will be through

- "partnership supplements", that we will describe a bit more detail in this presentation, and
- four (4) joint workshops mandatory for the awardees during Phase-I and II of the DARPA project.

slide 8
<Sankar Basu>

Some additional clarifications may be necessary in order to elaborate more on this particular joint/collaborative effort between NSF/DARPA.

First, it must be realized that the NSF-RTML and DARPA-RTML are two distinct programs. Thus, the solicitation, proposals submission, review criteria etc. for the two programs are different, and follow guidelines of the respective agencies.

Funding will not be co-mingled, i.e., no cofunding will be provided by DARPA for NSF projects, or by NSF for DARPA projects, in general.

However, since the two programs have closely related technical goals, we hope that it may provide opportunities for collaborations between the grantees through joint workshops and "partnership supplements", to be described shortly.

From a technical standpoint, NSF will strongly encourage, but will not insist on the two DARPA target application areas, i.e., ML for image processing systems and 5G applications mentioned earlier.

slide 9
<Sankar Basu>

As for more specific **technical goals** for collaboration, the 3 years of DARPA project will be divided into Phase I and Phase II - 18 months each, whereas NSF will provide three-year continuing grants.

The Phase I of DARPA project will produce a silicon compiler benchmarked using standard ML.

At the end of Phase I, the DARPA silicon compiler will be made available to NSF grantees for evaluation of their methods. However, this use of DARPA silicon compiler by the NSF grantees

is only optional, and if a more compelling evaluation method is available then the PIs are welcome to use them.

On the other hand,

Results of NSF awards after 18 months of the NSF project, will be made available to DARPA teams for implementation in Phase II.

NSF projects are expected to seek hardware implementation schemes integrable into silicon platforms, including in FPGAs or ASICs,

whereas

DARPA is more focused on design of ASICs in 14nm Silicon CMOS technologies by considering metrics, e.g., size, weight, power, latency, and platform adaptability.

Once again, NSF will encourage, but will not insist on the two DARPA target applications.

slide 10
<Sankar Basu>

This slide provides an assortment of standard ML techniques of interest, but proposal ideas need not be limited to these.

They include:
Feed forward (convolutional) neural networks, Recurrent networks; Neuroscience-inspired architectures, such as spike time-dependent neural nets, architectures derived from classical psychophysics and statistical mechanics; supervised, unsupervised or semi-supervised learning; or more modern techniques e.g., Generative Adversarial Networks (GAN).

The proposers may consult other approaches listed in both NSF solicitation and DARPA-BAA, which include, e.g., transfer learning, reinforcement learning, manifold learning etc.

Most importantly, however, for NSF-RTML routine implementations of known ML algorithms in existing hardware are not encouraged.

slide 11
<Sankar Basu>

Some more guidelines on relevant topics for this NSF program are:

- Hardware-software-algorithm cross-layer co-design is a key component of this program
- algorithms for all stages of training (e.g., active learning, incremental learning etc.) in real time are within scope
- metrics of performance, e.g., size, weight, power, latency and energy efficiency will be important

- Approximate algorithms will be given importance
- As mentioned earlier, ML for Distributed scenario may be considered
- For digital hardware, data and memory paths, e.g., in- or near-memory computations may be considered
- In addition to digital hardware, analog/mixed signal hardware will also be within scope of the NSF RTML program.

slide 12
<Sankar Basu>

We will now describe the NSF Solicitation Requirements.

slide 13
<Sankar Basu>

Information on award size for the NSF RTML program is as follows:
The anticipated NSF-funding amount is a total of $10M for 3 years.

There will be two types of awards in terms of their size:
- Small Awards: up to $500,000 for 3 years;
- Large Awards: up to $1,500,000 for 3 years.

We estimate a total of about number of 8 to 12 awards, both types taken together.

Approximately $1M of NSF program funds mentioned above will be reserved for "partnership supplements".

The anticipated type of award is "Continuing Grant" (as opposed to standard grants), which means that each year's funds will be given through annual increments.

As always, estimated program budget, number of awards, and average award size, or duration are subject to the availability of funds.

slide 14
<Sankar Basu>

The following **additional Budgetary Information** is specifically relevant to NSF-RTML program.

Proposals should budget for up to two project personnel to attend four joint NSF-DARPA workshops over the duration of the project (which is 36 months). The workshops are to be held in the Washington DC area.

Regardless of DARPA guidelines, as for most other NSF projects, inclusion of voluntarily committed cost sharing will be prohibited for NSF proposals.

slide 15
<Sankar Basu>

Organization Eligibility and Limit for NSF-RTML program are as follows:

There are no restrictions or limits on the number of proposals per organization, i.e, an organization e.g., an university, can submit any number of proposals.

U.S. academic institutions which perform research and with degree-granting education programs in disciplines normally supported by NSF are eligible to be the lead organization.

As in most NSF solicitations, academic institutions are defined as accredited two- and four-year Institutions of Higher Education having a campus in the US, acting on behalf of their faculty.

Involvement of an overseas branch of an US institution (including subawards/consultants), must justify why the project activities cannot be otherwise be performed within the US.

slide 16
<Sankar Basu>

PI Eligibility and Limitations for the NSF-RTML program are as follows:

An individual can participate as PI, co-PI, Senior Personnel, or Consultant on no more than two proposals submitted in response to this solicitation.

If more than two such proposals are received, then the first two proposals received will be accepted and the remainder will be **returned without review**.

It is important to remember that proposals submitted in response to this solicitation **may NOT *duplicate, or be substantially similar*** to other proposals concurrently under consideration by DARPA (similar proposals will be returned without review).

However, there are no restrictions on an **institution** as regards submitting an RTML proposal to both the DARPA-RTML program and NSF-RTML program. (other than, of course, duplicate/similar submissions mentioned above)

slide 17
<Sankar Basu>

More on NSF-RTML PI Eligibility and Limitations …

Principal Investigators must be at the faculty level as determined by the submitting organization i.e., universities.

DARPA Phase-I performers who do not move on to DARPA Phase-II will be eligible for NSF funding. Such PIs will need to partner with an existing NSF-RTML grantee on a **Large project**

for collaboration through **'partnership supplements'**. Such supplements will be given to large projects only.

On the other hand,

DARPA teams can include researchers from NSF-RTML awardees as part of their DARPA Phase 2 effort funded through 'partnership supplements' from DARPA.

slide 18
<Sankar Basu>

We now describe the NSF-RTML REVIEW CRITERIA.

The Generic NSF review criteria will apply to all NSF-RTML proposals.
These include National Science Board (NSB) approved Merit Review Criteria, including

- Intellectual Merit, and
- Broader Impacts

of the proposals under submission.

NSF Staff will also give careful consideration to:

- Integration of Research and Education
  and
- Integrating Diversity into NSF Programs, projects, and activities.

slide 19
<Sankar Basu>

There are additional NSF-RTML Solicitation Specific Review Criteria for this program.

All successful proposals need to address the following criteria carefully:

- Synergy in machine learning, software, algorithm, and hardware co-design to meet the real-time machine learning goals of this program.

- For Large proposals, the following additional review criteria will also be applied:
  1. strength of the Project Management and Collaboration Plan in the Project Description.
  2. strength of the Evaluation/Experimentation plan (e.g., in FPGA/ASIC) in the Project Description.

slide 20
<Sankar Basu>

Next, the timelines.

The program announcement was already made on March 8, 2019.

Proposals Submission Deadline for this NSF program is June 6, 2019, 5:00pm proposers local time. Proposals received after the deadline will not be considered.
We expect to review proposals in July 2019, and
NSF awards are expected to be made, pending availability of funds, by September 2019.
NSF projects are likely to begin in FY 2019 ending in Sept. 31, 2019.
A kick-off Meeting mandatory for PIs to attend, will be planned for Fall 2019 together with DARPA.

slide 21
<Sankar Basu>

We have received some questions from the community. We will go over them, and try to answer them first. Subsequently, we will take questions on-line.

slide 22
<Sankar Basu>
Qn 1 [Si-compiler]: The solicitation reads: "The DARPA Phase 1 objective is a RTML hardware silicon compiler, and the outcome will be made available by DARPA to the NSF awardees as an option to evaluate their proposed new RTML approaches. In the meantime, new techniques and results produced by NSF awardees during the first 18 months will be made available to DARPA project teams for them to implement in their Phase 2 efforts to explore novel ML architectures and circuits that will enable RTML."

Does this mean that NSF RTML projects cannot include a silicon compiler as this will be part of the DARP RTML focus? Or no such restriction exists?

Ans: No restriction exists. However, use of DARPA silicon compiler may set the stage for comparing techniques with other participants of the overall program.

slide 23
<Sankar Basu>
Qn 2 [FPGA]**:** We work in optimizing and hardware software codesign for neural models on FPGAs, with the intent of real-time performance. I noticed the following: "Proposers should be aware that routine implementations of existing AI/ML algorithms in standard hardware are not within scope of this program."

Does it mean (standard hardware) FPGAs are not relevant for this program?

Ans: A major goal of this joint program is to explore innovations in hardware-software-algorithms codesign for AI/ML. Use of standard techniques as only one aspect of the project, may or may not make it less compelling.

slide 24
<Sankar Basu>

Qn 3: [continuing grant]
The call mentions this is a continuing grant and not a standard grant. I wanted to make sure that this information is accurate. (The NSF guidelines state that the continuation of the grant is dependent on results.)

Ans: Continuing grant refers to the fact that annual installments of the 3 year grant is given every year. It is one of the standard NSF procedures used for many grants.

slide 25
<Sankar Basu>
Qn 4: [Industry/GOALI]

Is it okay to have an industrial partner, and whether having them as subcontractor is okay, in which case, they will receive funding through a subaward (what if it is a GOALI proposal?).

Ans: In this solicitation, proposals may only be submitted by Institutions of Higher Education (IHEs). If the proposal is submitted as a GOALI proposal with industrial partner, the submission guidelines of GOALI proposals in PAPPG should be followed, and the NSF funds are not permitted to be used to support the industrial research partner.

slide 26
<Sankar Basu>
Qn 5: [EDA techniques]
Will EDA and modeling techniques specialized for machine learning chip design be in the scope of this program?

Ans: ASICS may be considered.

slide 27
<Sankar Basu>
Qn 6: [applications]
Can applications of algorithms and hardware developed in the project be funded? Specifically, UAV applications, but where UAV infrastructure are supported by other funds.

Ans: Innovations in hardware-software-algorithm codesigns inspired by applications are within scope. Infrastructure is not supported by this NSF program.

slide 28
<Sankar Basu>

Some key websites for this program are listed in this last slide.

slide 29
<Sankar Basu>
Thank you.

Dr. Jenshan Lin and I and we will entertain online questions from the audience during the remaining time.