Experimental Demonstration of Conversion-Based SNNs with 1T1R Mott Neurons for Neuromorphic Inference

Xumeng Zhang[1,2,3], Zhongrui Wang[2], Wenhao Song[2], Rivu Midya[2], Ye Zhuo[2], Rui Wang[1,2,3], Mingyi Rao[2], Navnidhi K. Upadhyay[2], Qiangfei Xia[2], J. Joshua Yang[2]*, Qi Liu[1,3]*, and Ming Liu[1,3]*

[1]Key Laboratory of Microelectronics Devices and Integrated Technology, Institute of Microelectronics, Chinese Academy of Sciences, Beijing100029, China; [2]Department of Electrical and Computer Engineering, University of Massachusetts, Amherst MA 01003, USA; [3]University of Chinese Academy of Sciences, Beijing 100049, China.

*Email: jjyang@umass.edu; liuqi@ime.ac.cn; liuming@ime.ac.cn

Abstract- SNNs using the conversion-based approach could benefit the energy efficiency of inference and retain high accuracy of DLNs. However, transistor-based spiking neurons and synapses are not scalable and inefficient. In this work, a Mott neuron with 1T1R structure is designed to meet the requirement of the conversion-based approach, whose spiking rates dependence on voltage naturally implements the rectified linear unit (ReLU). Based on the 1T1R Mott neuron, we experimentally demonstrated a one-layer SNN (320 ×10), which consists of RRAM synaptic weight elements and Mott-type output neurons, for the first time. Attributes to the rectified linear voltage-rates relationship of the 1T1R neuron and its inherent stochasticity, *95.7%* converting accuracy of the neurons and *85.7%* recognition accuracy in *MNIST datasets* are obtained. At last, a neuron X-bar architecture is proposed for parallel multi-tasking and better system integration.

I. Introduction

Neuromorphic machines based on spiking neural networks (SNNs) have attracted great attention due to their high biomimetic and decent computational efficiency. However, in the algorithm level, the SNNs have two main restrictions: 1) lack of appropriate training algorithms; 2) insufficient spike-based datasets [1]. In the hardware level, neurons implemented by CMOS circuits usually consist of capacitors and more than 10 transistors, which limits the scalability [2]. While emerging technologies (e.g., Ferroelectric FET (Fe-FET) [3], Phase Change Memories (PCMs) [4], Spin Transfer Torque Magnetic RAMs (STT-MRAMs) [5], and Mott materials [6]) have been explored to emulate spiking neurons, experimental demonstrations of promising system performance are lacking.

In this work, we experimentally demonstrated a one-layer SNN (320 × 10) based on Mott-type spiking neurons and RRAM synapses for the first time (**Fig. 1(a)**). First, the conversion-based SNN is operated to keep the high performance of deep learning networks (DLNs), which correlates the spiking behavior of neurons to the ReLU activation function (**Fig. 1(b)**). Second, an analog input is retained in the first layer to alleviate the lack of spike-based datasets. Third, a 1T1R neuron is designed, whose firing rates dependence on gate voltage is proposed to implement the ReLU function. The 1T1R neurons are then connected with the trained memristor array (128 × 50). Experimental results show that the stochasticity of the Mott neurons make the accuracy (85.7%) nearly identical to that with software neurons (86%), even with 4.7% converting error of the Mott neurons. At the end, for parallel multi-tasking and better system integration, the X-bar architecture with 1T1R neurons is proposed. These results show that the 1T1R neurons are promising primitives to construct large-scale SNNs in the future.

II. Mott Neurons and Memristive Synapses

A. NbO_x Mott device and HfO_2 synapses

The Mott device in this work is fabricated with a-Si/NbO_x/TiN vertical structure (inset of Fig. 2). After forming, the device shows stable bi-directional Mott transitions (**Fig. 2**). **Fig. 3** shows the cumulative distribution of the threshold (V_{th}) and hold (V_{hold}) voltages in both biasing directions. Compact distribution of V_{th} ensures the stable firing of the neurons. The cycle-to-cycle fluctuations of the V_{th} and V_{hold} in both directions are shown in **Fig. 4**, cycle-independent distributions reveal the firing events of neurons does not interplay. The V_{th} and V_{hold} distributions of ten devices are shown (**Fig. 5**). Nearly identical distributions of V_{th} and V_{hold} with tiny fluctuations show that our devices could build high-precision neurons. A 128 × 64 1T1R memristor array serves as the synapses, in which a 128 × 50 sub-array will be used in this work [7].

B. 1T1R Mott neuron and training process

Given the above-mentioned characteristics of the NbO_x device, we designed a novel neuron circuit. The NbO_x device is serially connected to the drain of MOSFET in 1T1R structure (**Fig. 6**). The comparison of our neuron with previously reported Mott neurons is shown in **Table I**. In this work, we use the inherent parasitic capacitance for charge integration, which get rids of external capacitors. Channel resistance ($R_{channel}$) of the MOSFET serves as the integration resistor which is tunable under different input voltages. The equivalent circuit of the 1T1R neuron is shown in the right part of Fig. 6. The reconfigurable $R_{channel}$ under different input voltages induces different $\tau_{integration}$ ($\sim R_{channel}C_{parasitic}$) and then results in various spiking rates of the neurons. **Fig. 7** shows the output curves under different gate voltages of the 1T1R neuron. When the gate voltage is higher than 1.8 V, an abrupt switching is observed and the hysteresis window enlarges with increasing gate voltages. To complement the conversion-based SNN, analog voltages were applied on the gate and the output voltage was measured by oscilloscope, as shown in **Fig. 8**. **Fig. 9** shows the extracted statistical spiking

rates versus the gate-voltage relation of the neurons. A wider linear frequency range with about 500 kHz (more than 5 bits within 100 μs) is observed. This relation was mapped to ReLU function by shifting the origin to the starting voltage (inset of Fig. 9).

The LIF neural model is commonly used in SNNs. During training, the stochastic gradient descent algorithms are always performed on the real-valued membrane potential with the goal of having the correct output neuron fire more spikes [1]. The mathematic model of the LIF neurons is shown in **Fig. 10**. With a fixed time period, the membrane potential is proportional to the input currents. The same characteristic is observed in the 1T1R neuron (**Fig. 11**). Then the extracted membrane potential of the 1T1R neuron is used for training the network. The down-sampled binary MNIST datasets (20 × 16) is used as the input (**Fig. 12**). **Fig. 13** shows the accuracy vs. training epoch. The experimental accuracy was up to ~86%, which is a few percent lower than the result using software, owning to the down-sampled MNIST datasets and the non-ideal characteristics of memristor synapses.

III. Spiking Neural Network

C. Inference of the SNNs with Mott neurons

To test the system performance of the SNNs, ten neurons are serially connected to the memristor synaptic array, as shown in **Fig. 14**. In real operation, every input digit (20 × 16) is divided into five parts to feed the array separately, so does the training process. To represent negative weights, differential resistor pairs are used, so 640 inputs are required for each input digit pattern. The five output current vectors of each input digit are summed up to drive 1T1R neurons. Here a 1.81 V constant bias voltage is added to the gate voltage. **Fig. 15** shows the trained weight map. During testing, 100 μs read pulses of −0.2/0.2 V (differential paired voltages) amplitude represented pixel "1", and 0/0 V for pixel "0". **Fig. 16** shows the output of the "1" neuron evolves with 50 input digital patterns. Then we extracted the spiking rates of each neuron, as shown in **Fig. 17**. For each digit pattern, 10 output neurons with different spiking rates are observed, which means that the neurons can classify the input patterns. The classified digits with max frequency revealed by neurons "0" to "9" are labeled.

The recognition results are shown in **Fig. 18**. Using the same synaptic array, the ReLU software neurons have 1404 wrong predictions in classifying the entire 10000 testing data, achieving ~86% recognition accuracy (top panel of Fig. 18). The middle panel of Fig. 18 shows the converting errors of the Mott neurons compared with that of the software neurons, in which 247 labels are mismatched with the software neurons, corresponding to ~97.5% converting accuracy. However, in these 247 error labels, there are 85 labels (with red color) correct in comparison to the testing labels. This is because the dynamics of the neurons has a certain probability to lower the output frequency of the error neuron and elevate the output frequencies of the correct neuron. To test this hypothesis, we extracted the ten outputs corresponding to the corrected input digits, these outputs are from the software neurons (**Table II**). The outputs of the erroneous neurons are indeed slightly higher than the outputs of the correct neurons. The bottom panel of Fig. 18 shows the error labels of the Mott neurons compared to testing labels. For the 10000 test images, 85.7% recognition accuracy is obtained, which is nearly identical to that with software neuron (86%). These results demonstrate that our Mott neurons could convert analog signals to spiking rates decently and are able to enhance the recognition accuracy by its inherent dynamics.

D. Mott neurons in X-bar for multi-tasks

With the help of transistors in the neurons, a X-bar architecture with the 1T1R neurons, peripheral circuits, and integrated 3D memristor weight array is proposed for parallel multi-tasking and better system integration (**Fig. 19**(a)). In the 3D weight array, each row of multi-layers (or several rows) performs one task. In each inference step, the row is selectively activated by the MUX circuit. A simulation is performed with ten tasks in parallel. The different TIA outputs of the 10 tasks are shown in Fig. 19(b). **Fig. 20** shows the relative outputs when the TIA outputs of 10 tasks are fed the X-bar in parallel. For each task, the array can classify the TIA outputs independently. These results demonstrate that the proposed neuron X-bar architecture could be used for parallel multi-tasking applications.

IV. Conclusion

In this work, to construct high-performance hardware SNNs, a spiking Mott neuron with 1T1R structure is proposed to implement the ReLU activation function in DLNs. Based on this neuron and a 128 × 64 1T1R array, a one-layer conversion-based SNN (320 × 10) was experimentally demonstrated. The SNN achieved 95.7% converting accuracy using the Mott neurons and nearly software neuron-like recognition accuracy (85.7%) in classifying the MNIST datasets. At last, a neuron X-bar architecture based on 1T1R neurons was proposed for muli-tasking applications. The results demonstrated that the Mott neurons with 1T1R structure could serve as a promising candidate for fabricating hardware SNNs chip in the future.

Acknowledgment

This work was supported by the National High Technology Research Development Program under Grant No. 2017YFB0405600, the National Natural Science Foundation of China under Grant Nos. 61825404, 61732020, 61821091, 61888102, the Strategic Priority Research Program of the Chinese Academy of Sciences under Grant No.XDPB12.

References

[1] B. Rueckauer *et al.*, arXiv preprint *arXiv:1612.04052* (2016).
[2] S. Kim, *et al*, *IEDM Tech. Dig.,* pp. 17.1.1-17.1.4, (2015).
[3] Z. Wang, *et al., IEDM Tech. Dig.,* pp. 13.3.1-13.3.4, (2018).
[4] T. Tuma, *et al., Nat. Nanotechnol*. **11**, 693-699 (2016).
[5] T. Talatchian, *et al., IEDM Tech. Dig.,* pp. 27.4.1-27.4.4, (2018).
[6] J. Lin, *et al., IEDM Tech. Dig.,* pp. 34.5.1-34.5.4, (2016).
[7] Z. Wang, et al., *Nat. Electron*. 2, 115-124, (2019).
[8] W. Yi, *et al., Nat. Commun.* **9**, p. 4661 (2018).
[9] M. Jerry, *et al., Sym. VLSI Tech. Dig.*, T186-T187 (2017).
[10] Q. Luo, *et al., Sym. VLSI Tech. Dig.*, T236-T237 (2019).

Conversion-based SNNs and ReLU function

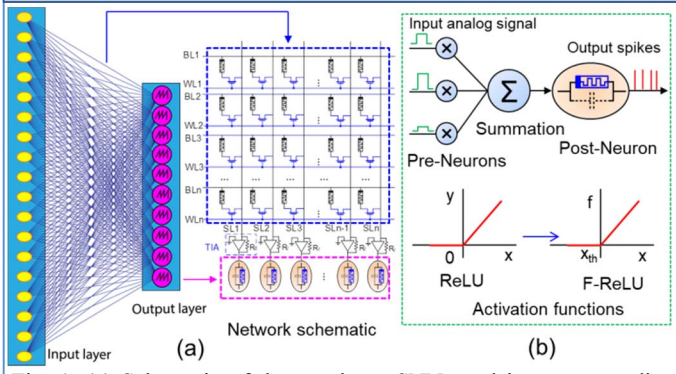


Fig. 1. (a) Schematic of the one-layer SNNs and its corresponding parts in hardware. (b) The ReLU function in DLNs and the F-ReLU activation function in SNNs, analog inputs are operated.

Comparison of different Mott spiking neurons

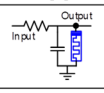
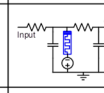
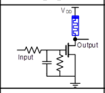
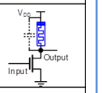
Mott Spiking Neurons				
References	[6]	[8]	[9]	Our work
Circuit schematic				
Model	LIF	H-H	LIF	LIF
Input signal	Analog/spikes	Analog/spikes	Spikes	Analog/spikes
Function	#	#	Sigmoid	ReLU
Integration	Low	Low	Low	High(X-bar)
System	Simulation	#	Simulation	Experiment

Table I. Comparison between reported neurons and our work. The neuron is designed base on the LIF neuron model and supports both analog (for conversion-based SNNs) and spikes inputs signal. The ReLU function is introduced into the neuron. The 1T1R structure supports X-bar integration.

Characteristics of the NbO_x Mott device

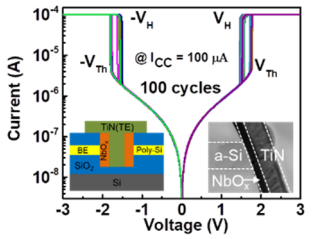


Fig. 2. 100 cycles of DC sweeps of the NbO_x device, inset: device structure the TEM image.

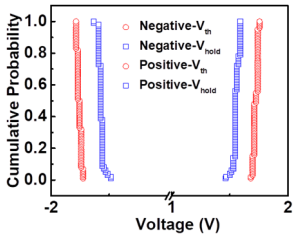


Fig. 3. Threshold and hold voltages distributions in both directions.

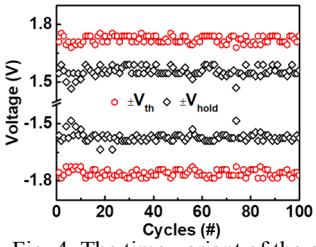


Fig. 4. The time-variant of the cycle to cycle fluctuations. Tested to support > 10^{12} events [10].

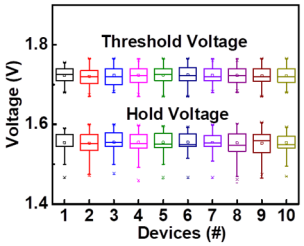


Fig. 5. The V_{th} and V_{hold} distributions on positive direction of ten devices.

1T1R spiking neuron and training process

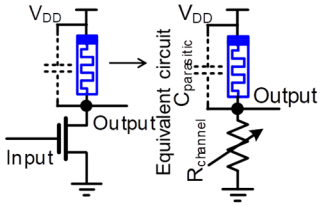


Fig. 6. The 1T1R neuron schematic and its equivalent circuit.

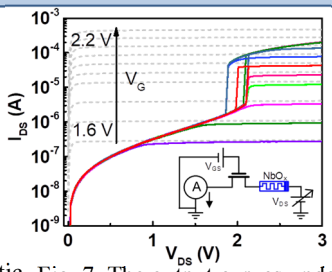


Fig. 7. The output curves under different gate voltages of the 1T1R neuron.

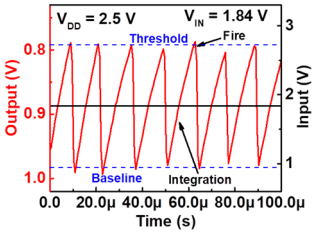


Fig. 8. The integration and fire behavior of the neuron under a 1.84 V input voltage.

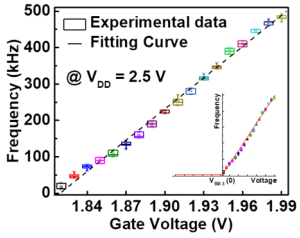


Fig. 9. Extracted statistical spiking rates-gate-voltage relation of the neurons. Inset: The F-ReLU function.

$$\tau_m \frac{du}{dt} = -u(t) + RI(t) + u_{rest}$$
$$if\ u(t) \geq u_{th}$$
$$u(t) \rightarrow u_{rest}$$

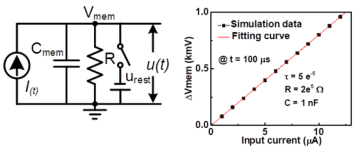
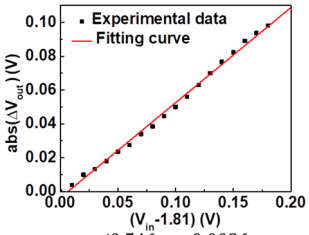


Fig. 10. The LIF neuron model and the changed membrane potential under different input currents at 100 µs.



$$abs(\Delta Vout) = \begin{cases} 0.546x - 0.0036, & x > 0 \\ 0, & x < 0 \end{cases}; \quad x = V_{in} - 1.81$$

Fig. 11. The extracted changed membrane potential of the 1T1R neuron, and the fitted ReLU function for training.

Fig. 12. Down-sampled binary MNIST images (20 × 16).

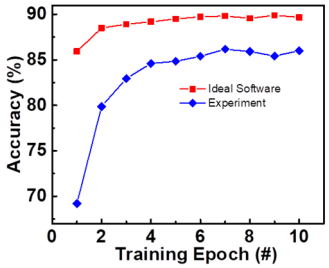


Fig. 13. Accuracy vs. training epoch. One epoch is 60,000 training images. Test is done by another set of 10,000 testing images.

SNNs with 1T1R neurons and memristive synapses

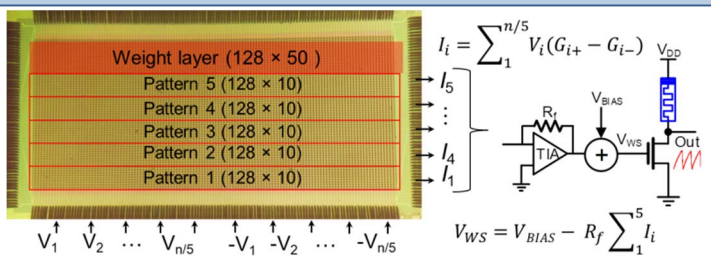


Fig. 14. The hardware schematic of the SNNs. Each input digit (20 × 16) is divided into five parts to load into the array separately. A differential resistor pair is used to represent a negative weight, so 640 inputs are required for each input digit pattern. Then 5 outputs of each input digit (I1-I5) is added together and the converted into voltage by a TIA to serve as the input of 1T1R neuron.

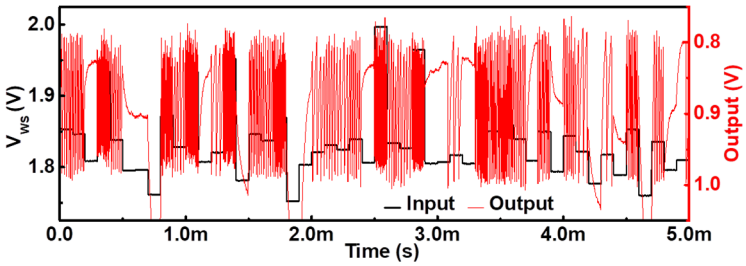


Fig. 16. The output of the "1" neuron evolves with 50 input digital images. The neuron spikes in a higher frequency under higher WS voltages.

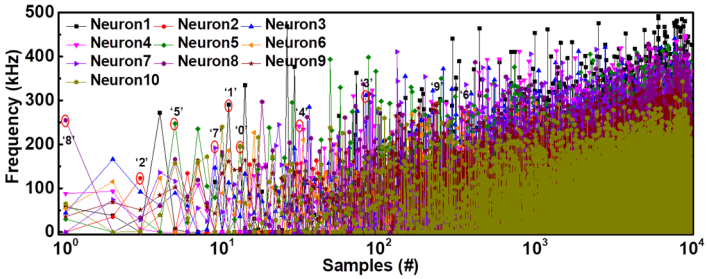


Fig. 17. The extracted the spiking rates from each neuron. Each digit pattern has 10 output spiking rates. The classified digits with max frequency revealed by neurons "0" to "9" are labeled

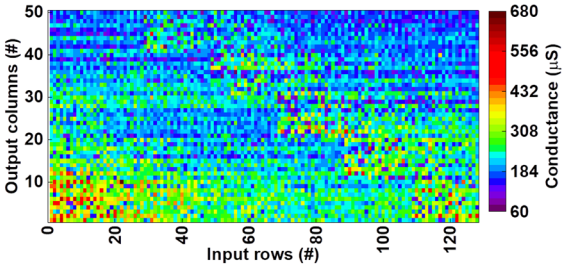


Fig.15. The trained weight map of the one-layer SNN.

Table II. Ten outputs correspond to the corrected input digits. **Red color indicates the output neuron with max value, blue color indicates the output neuron with secondary value.

	Image 1	Image 2	Image 3	Image 4	Image 5	Image 6	Image 7	Image 8	Image 9	Image1 0
Neuron1	0.132	0.089	0	0.006	0.133	0	0.283	0	0	0.283
Neuron2	0.276	0	0	0	0.229	0	0	0.052	0.218	0
Neuron3	0.268	0	0.481	0.218	0.129	0.225	0.346	0	0.385	0.346
Neuron4	0.262	0.247	0.479	0	0.267	0	0	0.239	0.388	0
Neuron5	0	0.176	0	0.439	0	0.369	0.015	0.386	0	0.015
Neuron6	0.214	0.349	0.165	0	0.266	0.133	0.119	0.161	0.117	0.119
Neuron7	0.046	0.047	0.005	0.34	0	0.373	0.347	0.14	0.014	0.347
Neuron8	0	0.237	0.208	0.243	0.165	0.171	0	0.358	0.131	0
Neuron9	0.241	0.352	0	0.345	0.2	0.27	0.097	0.176	0.4	0.097
Neuron0	0	0.318	0	0.436	0.094	0.155	0	0.381	0.044	0

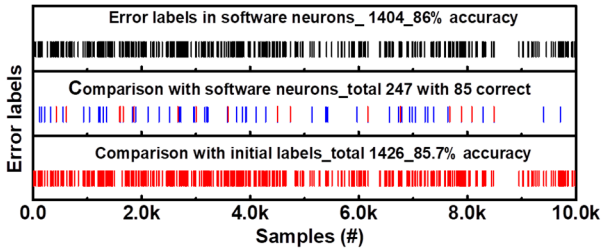


Fig. 18. Top panel: error labels from the software neurons. Middle panel: error labels of the Mott neurons that compared to the labels of the software neurons, red labels indicate the corrected labels by the 1T1R neurons. Bottom panel: error labels of the Mott neurons compared to testing labels.

Neuron X-bar for parallel multi-tasking and system integration

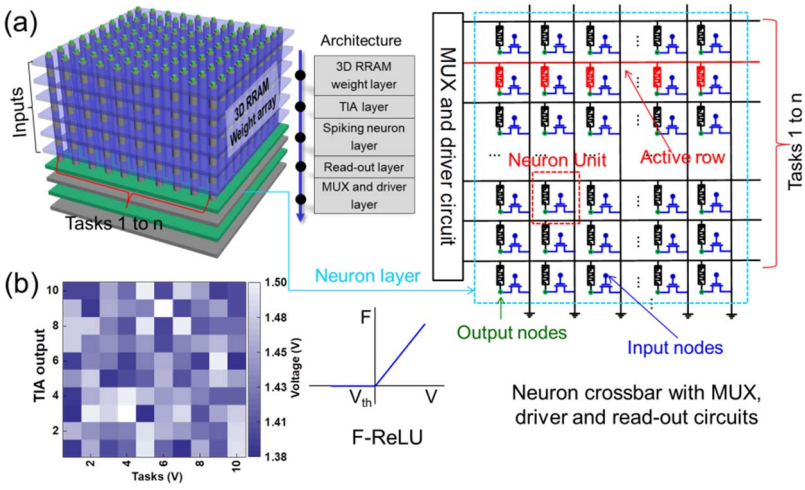


Fig. 19. (a) The proposed neuron X-bar architecture and its peripheral circuits. A 3D memristor weight array is designed for weight storage. (b) The TIA outputs of 10 tasks.

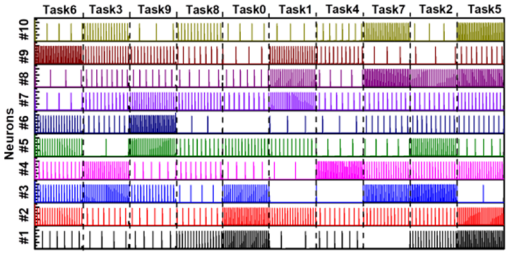


Fig. 20. The relative outputs when the TIA outputs of 10 tasks are loaded into the X-bar in parallel.