

[0008] The subject matter described in this specification can be implemented in particular embodiments so as to realize one or more of the following advantages. Shifting the sparse-to-dense data loading task from the central processing unit to specialized matrix processors increases the computation bandwidth of the central processing unit and decreases the processing cost of the system. The matrix processors can be arranged to be near the memory where the data is stored, and can lower the latency for loading data. By using specialized matrix processors, the use of processors that are specialized for dense linear algebra to fetch sparse data can be avoided. By having multiple units per-controller-channel, the number of simultaneous transactions that are being serviced at a time may be parallelized, and transactions may be processed immediately without waiting for a previous transaction to complete.

[0009] Other implementations of this and other aspects include corresponding systems, apparatus, and computer programs, configured to perform the actions of the methods, encoded on computer storage devices. A system of one or more computers can be so configured by virtue of software, firmware, hardware, or a combination of them installed on the system that in operation cause the system to perform the actions. One or more computer programs can be so configured by virtue of having instructions that, when executed by data processing apparatus, cause the apparatus to perform the actions.

According to one aspect, there is provided a system comprising a hardware sparse-dense transform unit for transforming sparse elements into a dense matrix, the system comprising: a sparse element access unit in the sparse-dense transform unit and for accessing multiple sparse elements of a dense matrix, wherein the sparse element access unit includes: a request identification unit configured to: receive, over a node network, a request for particular sparse elements stored in one or more data shards; determine that the sparse element access unit in which the request identification unit is located is assigned to handle a subset of the particular sparse elements; and in response to determining that the sparse element access unit in which the request identification unit is located is assigned to handle the subset of the particular sparse elements, generate an indication for accessing the subset of the particular sparse elements; at least one processor for fetching, based on the indication, one or more sparse elements of the subset of the particular sparse

elements; and a concatenation unit configured to: generate an output dense matrix based on a transformation that is applied to at least the one or more sparse elements.

According to another aspect, there is provided a method for transforming sparse elements into a dense matrix, comprising: receiving, by a sparse element access unit and from an external source, a request for particular sparse elements stored in one or more data shards, where the sparse element access unit is located in a sparse-dense transform unit; determining, by a request identification unit located in the sparse element access unit and based on the request, that the sparse element access unit in which the request identification unit is located is assigned to handle a subset of the particular sparse elements; in response to determining that the sparse element access unit in which the request identification unit is located is assigned to handle the subset of the particular sparse elements, generating, by the request identification unit, an indication for accessing the subset of the particular sparse elements; fetching, based on the indication and by at least one processor of the sparse element access unit, one or more sparse elements of the subset of the particular sparse elements; and generating, by a concatenation unit located in the sparse element access unit, an output dense matrix based on a transformation that is applied to at least the one or more sparse elements.

According to yet another aspect, there is provided a system comprising a hardware sparse-dense transform unit for transforming sparse elements to a dense matrix, the system comprising: one or more processors configured to send a request for an output matrix based on particular sparse elements that are stored in one or more data shards; a sparse element access unit in the sparse-dense transform unit and for accessing the particular sparse elements that are stored in the one or more data shards, the sparse element access unit comprising: a request identification unit configured to: receive, over a node network, the request for the output matrix based on the particular sparse elements that are stored in one or more data shards; determine that the sparse element access unit in which the request identification unit is located is assigned to handle a subset of the particular sparse elements; and in response to determining that the sparse element access unit in which the request identification unit is located is assigned to handle the subset of the particular sparse elements, generate an indication for accessing the subset of the particular sparse

elements; at least one processor for fetching, based on the indication, one or more sparse elements of the subset of the particular sparse elements; and a concatenation unit configured to: generate an output dense matrix based on a transformation that is applied to at least the one or more sparse elements.

[0010] The details of one or more implementations of the subject matter described in this specification are set forth in the accompanying drawings and the description below. Other potential features, aspects, and advantages of the subject matter will become apparent from the description, the drawings, and the claims.

BRIEF DESCRIPTION OF THE DRAWINGS

[0011] FIG. 1 is a block diagram of an example computation system.

[0012] FIGS. 2A-2D illustrate an example sparse-dense transform unit.

[0013] FIGS. 3A-3B illustrate an example sparse element access unit.

[0014] FIG. 4 is a flow diagram that illustrates an example of a process for generating a dense matrix.

CLAIMS

1. A system comprising a hardware sparse-dense transform unit for transforming sparse elements into a dense matrix, the system comprising:

a sparse element access unit in the sparse-dense transform unit and for accessing multiple sparse elements of a dense matrix, wherein the sparse element access unit includes:

a request identification unit configured to:

receive, over a node network, a request for particular sparse elements stored in one or more data shards;

determine that the sparse element access unit in which the request identification unit is located is assigned to handle a subset of the particular sparse elements; and

in response to determining that the sparse element access unit in which the request identification unit is located is assigned to handle the subset of the particular sparse elements, generate an indication for accessing the subset of the particular sparse elements;

at least one processor for fetching, based on the indication, one or more sparse elements of the subset of the particular sparse elements; and

a concatenation unit configured to:

generate an output dense matrix based on a transformation that is applied to at least the one or more sparse elements.

2. The system of claim 1, wherein determining that the sparse element access unit in which the request identification unit is located is assigned to handle the subset of the particular sparse elements comprises determining that the sparse element access unit in which the request identification unit is located is assigned to handle a subset of the particular sparse elements based on a lookup table.

3. The system of claim 1, the sparse element access unit further comprising:
a sparse reduce unit configured to:

receive a first matrix that includes a first sparse element from the at least one processor, the first matrix having a first dimension; and
generate a second matrix that includes the first sparse element, the second matrix having a second dimension that is smaller than the first dimension, wherein the concatenation unit is further configured to:
receive the second matrix,
wherein generating the output dense matrix further comprises generating the output dense matrix based on the second matrix.

4. The system of claim 1, wherein the concatenation unit is configured to:
receive a first sparse element at a first point of time;
receive a second sparse element at a second, different, point of time; and
determine an order of the first sparse element and the second sparse element for the output dense matrix,
wherein generating the output dense matrix further comprises generating the output dense matrix based on the order of the first sparse element and the second sparse element.

5. The system of claim 1, the sparse element access unit further comprising:
a compress/decompress unit configured to:
compress the output dense matrix to generate a compressed output dense matrix; and
provide the compressed output dense matrix to a node network.

6. The system of claim 5,
wherein the concatenation unit is further configured to receive a first dense matrix representing a dense matrix sent over a node network, and
wherein generating the output dense matrix further comprises generating the output dense matrix based on the first dense matrix, a first sparse element, and a second sparse element.

7. The system of claim 6, wherein the compress/decompress unit is configured to decompress a compressed first dense matrix to generate the first dense matrix.

8. The system of claim 1, wherein one or more sparse elements of the particular sparse elements are multi-dimensional matrices, and wherein the output dense matrix is a vector.

9. A method for transforming sparse elements into a dense matrix, comprising:
receiving, by a sparse element access unit and from an external source, a request for particular sparse elements stored in one or more data shards, where the sparse element access unit is located in a sparse-dense transform unit;

determining, by a request identification unit located in the sparse element access unit and based on the request, that the sparse element access unit in which the request identification unit is located is assigned to handle a subset of the particular sparse elements;

in response to determining that the sparse element access unit in which the request identification unit is located is assigned to handle the subset of the particular sparse elements, generating, by the request identification unit, an indication for accessing the subset of the particular sparse elements;

fetching, based on the indication and by at least one processor of the sparse element access unit, one or more sparse elements of the subset of the particular sparse elements; and

generating, by a concatenation unit located in the sparse element access unit, an output dense matrix based on a transformation that is applied to at least the one or more sparse elements.

10. The method of claim 9, further comprising:

receiving, by a sparse reduce unit, a first matrix that includes a first sparse element from the at least one processor, the first matrix having a first dimension;

generating, by the sparse reduce unit, a second matrix that includes the first sparse element, the second matrix having a second dimension that is smaller than the first dimension; and

receiving, by the concatenation unit, the second matrix,

wherein generating the output dense matrix further comprises generating the output dense matrix based on the second matrix.

11. The method of claim 9, wherein generating the output dense matrix further comprises:

receiving a first sparse element at a first point of time;

receiving a second sparse element at a second, different, point of time;

determining an order of the first sparse element and the second sparse element for the output dense matrix; and

generating the output dense matrix based on the order of the first sparse element and the second sparse element.

12. The method of claim 9, further comprising:

compressing the output dense matrix to generate a compressed output dense matrix; and

providing the compressed output dense matrix to a node network.

13. The method of claim 9, wherein generating the output dense matrix further comprises:

receiving a first dense matrix representing a dense matrix sent over a node network; and

generating the output dense matrix based on the first dense matrix, a first sparse element, and a second sparse element.

14. The method of claim 10, wherein determining that the sparse element access unit is assigned to handle the subset of the particular sparse elements comprises

determining that the sparse element access unit is assigned to handle a subset of the particular sparse elements based on a lookup table.

15. A system comprising a hardware sparse-dense transform unit for transforming sparse elements to a dense matrix, the system comprising:

- one or more processors configured to send a request for an output matrix based on particular sparse elements that are stored in one or more data shards;

- a sparse element access unit in the sparse-dense transform unit and for accessing the particular sparse elements that are stored in the one or more data shards, the sparse element access unit comprising:

- a request identification unit configured to:

- receive, over a node network, the request for the output matrix based on the particular sparse elements that are stored in one or more data shards;

- determine that the sparse element access unit in which the request identification unit is located is assigned to handle a subset of the particular sparse elements; and

- in response to determining that the sparse element access unit in which the request identification unit is located is assigned to handle the subset of the particular sparse elements, generate an indication for accessing the subset of the particular sparse elements;

- at least one processor for fetching, based on the indication, one or more sparse elements of the subset of the particular sparse elements; and

- a concatenation unit configured to:

- generate an output dense matrix based on a transformation that is applied to at least the one or more sparse elements.

16. The system of claim 15, the sparse element access unit further comprising:

- a sparse reduce unit configured to:

- receive a first matrix that includes a first sparse element from the at least one processor, the first matrix having a first dimension; and

generate a second matrix that includes the first sparse element, the second matrix having a second dimension that is smaller than the first dimension, wherein the concatenation unit is further configured to:

receive the second matrix,

wherein generating the output dense matrix further comprises generating the output dense matrix based on the second matrix.

17. The system of claim 15, the sparse element access unit further comprising:

a compress/decompress unit configured to:

compress the output dense matrix to generate a compressed output dense matrix; and provide the compressed output dense matrix to a node network.