DEUTSCHES ELEKTRONEN-SYNCHROTRON DESY

DESY DV-71/2
Dezember 1971

HYBRID

Ein Programm zur statistischen Analyse
von numerischen Daten

von

H. Butenschön

2 HAMBURG 52 · NOTKESTIEG 1

To be sure that your preprints are promptly included in the HIGH ENERGY PHYSICS INDEX, send them to the following address (if possible by air mail):

DESY
Bibliothek
2 Hamburg 52
Notkestieg 1
Germany

Abstract:

HYBRID is a program for analysing numerical data.
Its features are:

- The program execution time is shorter than conventional programs by approx. a factor of 10.
- Unlimited number of one-dimensional and two-dimensional plots in a single data-pass.
- Requires only small core-memory (~ 80k Bytes).
- The programmer of HYBRID needs only a little knowledge of FORTRAN.

INHALT:

H Y B R I D

Zusammenfassung

HYBRID dient zur statistischen Analyse von numerischen Daten. Es bietet gegenüber anderen Programmen dieser Art folgende Vorteile:

- HYBRID ist um eine Größenordnung schneller als konventionelle Programme
- " Unbegrenzte" Anzahl von ein- und zweidimensionalen Verteilungen bei einmaliger Verarbeitung der zu analysierenden Daten
- Geringer Kernspeicherbedarf (~80k Bytes à 8 bits bei IBM/360)
- Zum Betrieb des Programms sind lediglich geringe FORTRAN-Kenntnisse nötig.

1. EINLEITUNG

1.1 Zweck des HYBRID-Systems

HYBRID[1] ist ein FORTRAN-Programm zur Analyse von numerischen Daten. Diese Daten können auf Datenträgern vorhanden sein oder im HYBRID vom Benutzer selbst erzeugt werden. HYBRID stellt auf Wunsch durch einfache FORTRAN-Befehle von den Daten beliebige ein- und zweidimensionale Verteilungen her. Jede andere beliebige Datenmanipulation (z. B. Bildung von Mittelwerten und Varianzen mit vorhandenen Unterprogrammen) ist möglich, soweit sie sich mit FORTRAN programmieren läßt.

1.2 Vergleich mit anderen Analyse-Programmen

Im Unterschied zu anderen Datenanalyse-Programmen, die Datenmanipulation fast nur mit Hilfe einer Pseudosprache auf Datenkarten betreiben (z.B. SUMX[2]), braucht der Anwender bei HYBRID nur einige Grundkenntnisse in FORTRAN. Diese scheinbare Beschränkung führt zu einer weitgehenden Flexibilität des Programmsystems.

Als wesentliche Limitierung aller bisherigen Datenanalyse-Programme hat sich der beschränkte Speicherplatz herausgestellt, der für zweidimensionale Darstellungen zur Verfügung steht. Für ein Raster von nur 120 x 120 Fächern ist ein Speicherplatz von 60k Bytes erforderlich. Selbst eine Maschine der 3. Generation kann bei sinnvoller Performance nur etwa Platz für 3-10 solcher Darstellungen bieten. Das ist jedoch in den meisten Anwendungen zu wenig. Die Daten müssen also mehrfach verarbeitet werden. Dieses Verfahren findet seine Grenze jedoch in der zu verarbeitenden Datenmenge, und diese Grenze wird z.B. von den heutigen Experimenten der Hochenergiephysik[3] bereits wesentlich überschritten.

---

[1] Programmbeschreibungen: DESY 66/29, 1966
DESY-R1/1 , 1969
DESY-R2 , 1971 in Vorbereitung

[2] CERN COMPUTER 6000 SERIES LWV Y200

[3] Datenumfang des DESY γd-Experimentes ist $\sim 3*10^8$ Bytes

Ein effektives Verfahren wurde mit dem Programm HYBRID realisiert. Hier besteht weder eine Beschränkung in der Anzahl der zweidimensionalen noch der eindimensionalen Verteilungen bei nur einmaliger Bearbeitung der darzustellenden Daten. Die Verwirklichung von HYBRID wurde möglich durch die Verflechtung von wissenschaftlicher Anwendung mit kommerziellen Methoden. Es sind nämlich in der kommerziellen Anwendung seit langem effektive Sortierprogramme bekannt. Durch die Einbeziehung dieser Programme in HYBRID ist es möglich, trotz unlimitierter Anzahl von zweidimensionalen Darstellungen, Zeiteinsparungen um mehr als eine Größenordnung gegenüber bisherigen Programmen zu erzielen.

1.3 Entwicklung von HYBRID

HYBRID ist aus dem FORTRAN-Programmsystem ULTRAN[4] entstanden. Die Gemeinsamkeit mit ULTRAN besteht im FORTRAN-Kern, der dem Benutzer schon eine weitgehende Flexibilität erlaubte. ULTRAN hatte somit auch schon den Vorteil, daß der Benutzer nicht erst eine Pseudosprache erlernen mußte, um seine Wünsche dem Programm mitzuteilen. Diese Bequemlichkeit sowie die wachsenden Anforderungen der DESY-Experimentiergruppen führten jedoch bald dazu, daß ein beachtlicher Teil der gesamten Rechenmaschinenzeit durch ULTRAN-Programme verbraucht wurde. Durch das neue Konzept von HYBRID konnte ein großer Anteil an Rechenzeit eingespart werden; manche Projekte wurden erst denkbar durch das Vorhandensein von HYBRID.

---

[4] ULTRAN - a kinematical program for reduction of bubble chamber data, V.Blobel, H.Butenschön, P.v.Handel, P.K.Schilling, Hamburg 1964

2. Arbeitsweise von HYBRID

2.1 Vergleich mit anderen Programmen

Das vereinfachte Flußdiagramm eines konventionellen Analyse-Programms zeigt die Fig. 1 :

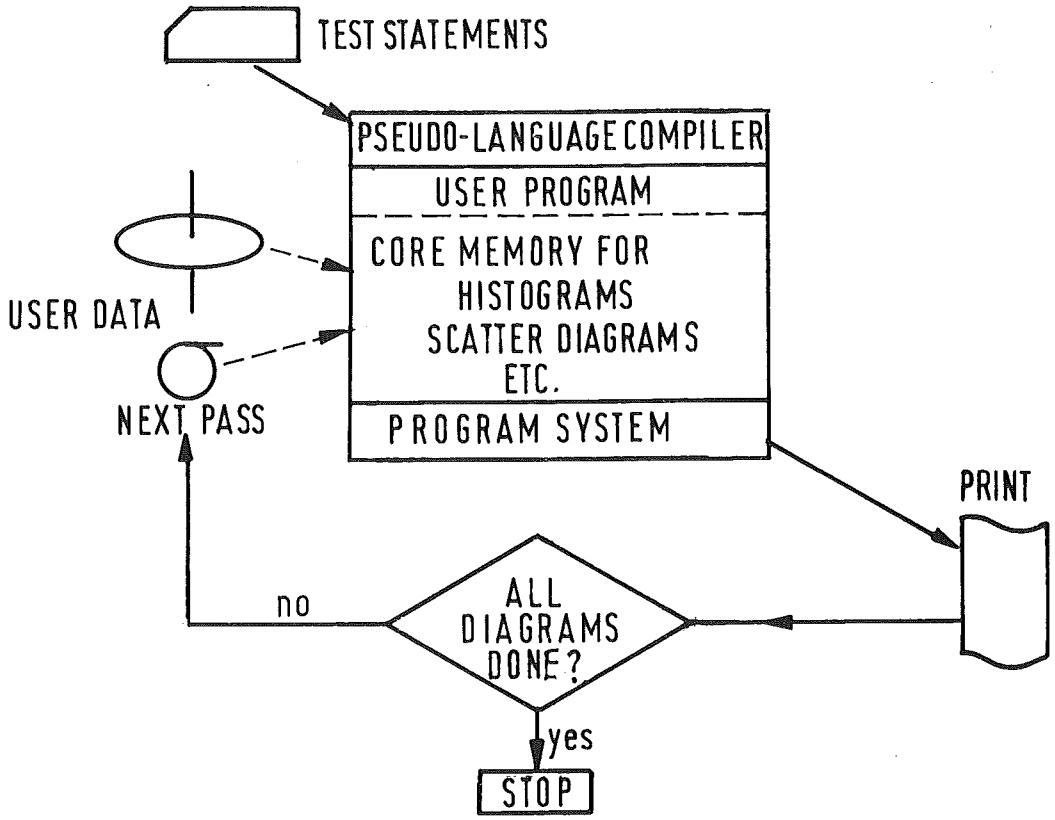

Fig. 1 Fluß-Diagramm eines konventionellen Analyse-Programms

Ein konventionelles Programm richtet zunächst den Kernspeicherplatz für die ein- bzw. zweidimensionalen Darstellungen ein. Jedem Kernspeicherplatz wird die Nummer der Darstellungen und die Position in dieser zugeordnet. Jedes darzustellende Datum wird nun auf dem entsprechenden Speicherplatz akkumuliert. Am Ende

der Eingabedaten kann man direkt den Inhalt der Speicherplätze ausdrucken und ist fertig, wenn die Anzahl der verlangten Darstellungen gering genug war. Sonst müssen die Daten mehrmals gelesen werden.

HYBRID geht anders vor (Fig. 2):

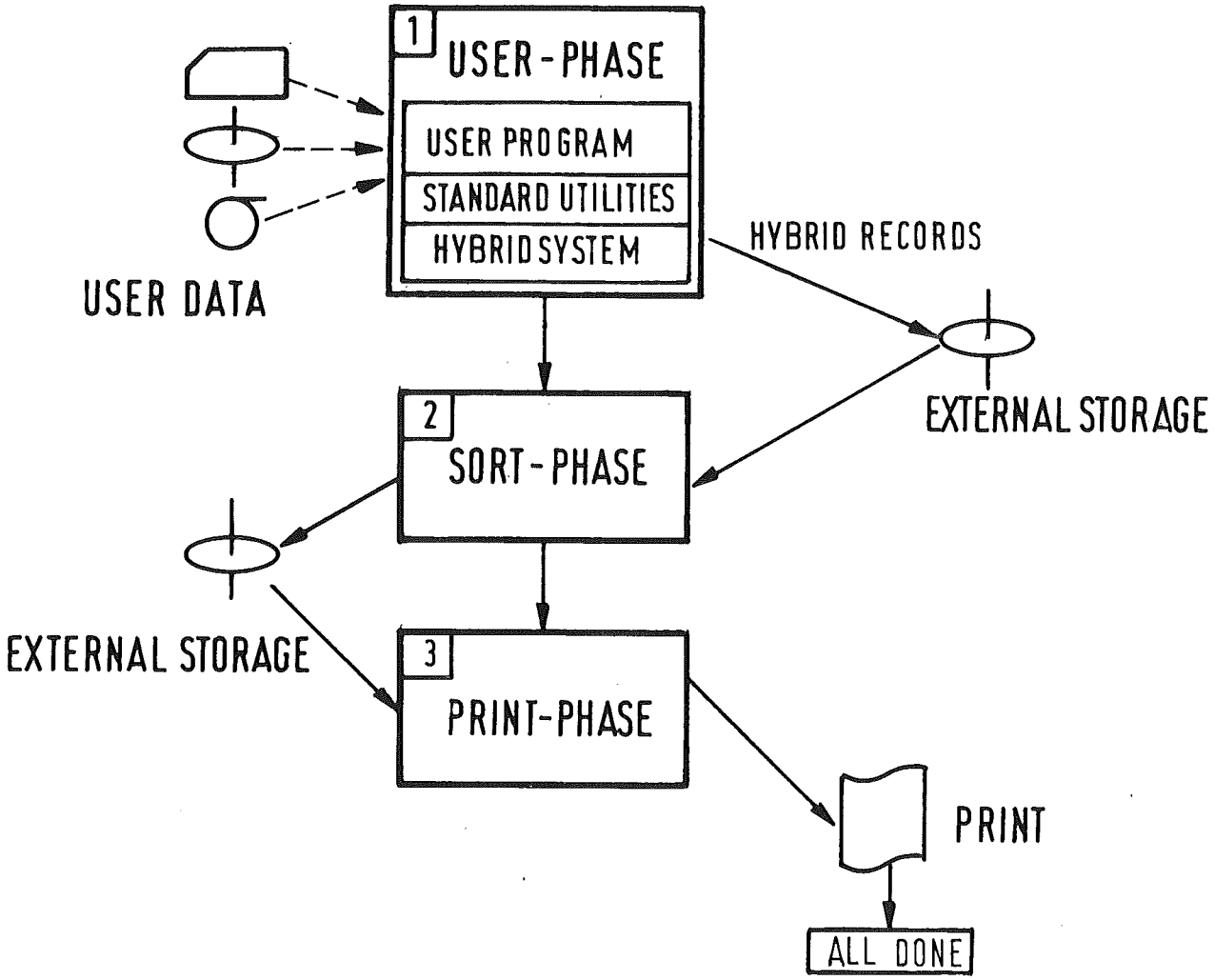

Fig.2 HYBRID - Flußdiagramm

Jedem darzustellenden Wert (oder Wertepaar) wird explizit die Nummer der Darstellung sowie ein Gewicht angeheftet.
Diese Information wird als HYBRID-Record (Fig. 3)

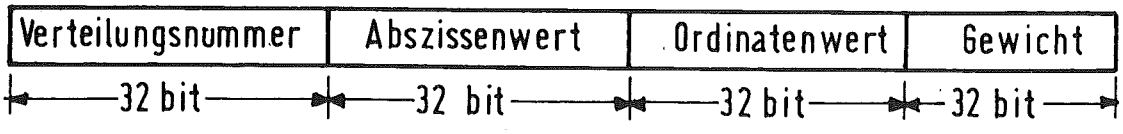

Fig.3 HYBRID-Record

von 128 bit Länge auf einen speziellen Datenträger geschrieben.

Es wird also im Gegensatz zu konventionellen Methoden kein Kernspeicherplatz für die Speicherung der Verteilung verbraucht; die Anzahl der Verteilungen ist damit theoretisch unlimitiert[5]. Nach der Verarbeitung der Daten besitzt man nun aber noch keine direkt ausdruckbaren Verteilungen. Dazu müssen erst die HYBRID-Records auf dem Zwischen-Datenträger nach der Verteilungsnummer sortiert werden. Diese Aufgabe wird von kommerziellen Sortiertechniken in erstaunlich kurzer Zeit bewältigt [6]. Nach Sortierung der Zwischendaten wird für die zweidimensionalen Darstellungen ein Bereich von n*m Speicherplätzen[7] eingerichtet. Der Datensatz wird nun so lange gelesen und verarbeitet, bis sich die Verteilungsnummer ändert. Dann wird die fertige Darstellung ausgedruckt, und der Speicherplatz steht zur Akkumulation der nächsten Verteilung zur Verfügung. Die Erfahrung mit HYBRID zeigt (es wird bei DESY seit etwa 5 Jahren benutzt), daß ~10% der Gesamt-CPU-Zeit eines Jobs für die Sortierung benötigt wird. D.h., daß nur bei einer geringen Anzahl von Darstellungen (~ 3-10) ein zeitlicher Mehraufwand gegenüber konventionellen Programmen von ~10% erforderlich ist. Gerade bei größeren Datenmengen wird aber der Wunsch nach einer größeren Anzahl von Verteilungen dringlich.

---

[5] Grenzen werden durch die spezielle Installation gegeben. Sie liegt für die Implementierung auf der DESY/360 bei ~ 10^4 Verteilungen.

[6] z.B. 10^5 Records in 40 sec CPU-TIME bei Verwendung von IBM-SORT.

[7] In HYBRID ist n=m=120; eine größere Auflösung ist bei der DESY/360 nicht notwendig, da der Ausdruck mit einem Schnelldrucker (≤ 132 Spalten) vorgenommen wird.

Die Fig. 4 zeigt schematisch einen CPU-TIME Vergleich mit konventionellen Programmen:

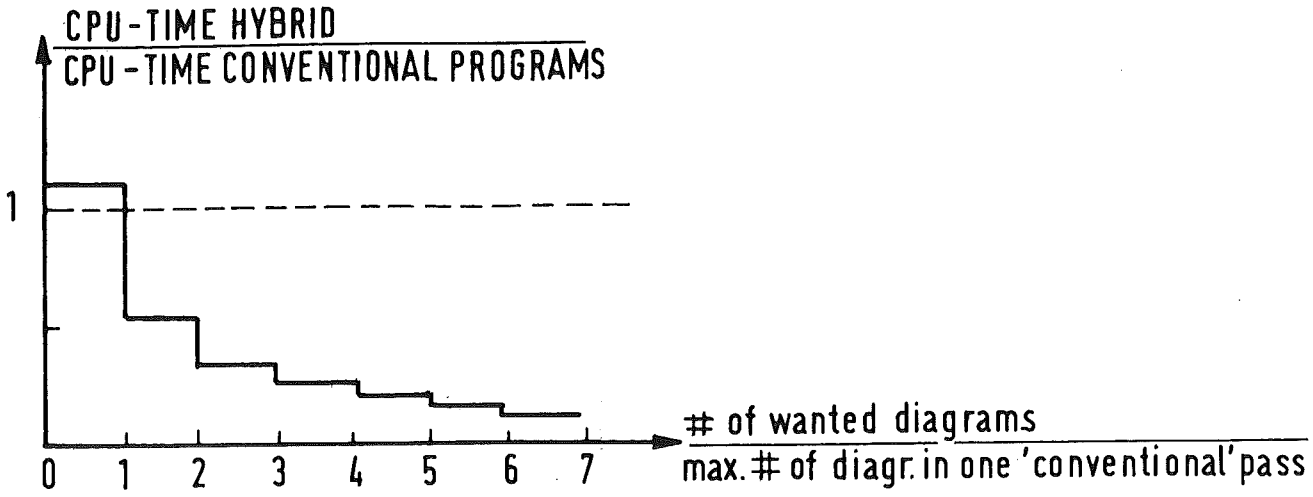

Fig.4 CPU-TIME-Vergleich

Die Sprünge im Diagramm sind dadurch verursacht, daß ein konventionelles Programm bei Überschreitung einer gewissen Zahl von Verteilungen die Daten nochmals bearbeiten muß.

Ein REAL-TIME-Vergleich ist sehr stark von der Rechenmaschinenkonfiguration abhängig (von Anzahl und Art der verfügbaren externen Speicher u.s.w.), so daß an dieser Stelle kein Vergleich gemacht werden kann.

Ein zweiter wesentlicher Vorteil besteht darin, daß der ganze Speicherplatz, den sonst zweidimensionale Verteilungen verbrauchen, dem Benutzer zugute kommt. Damit spart er in vielen Fällen die Vorarbeit, voranalysierte Daten zunächst auf einen Zwischendatenträger zu schreiben, um sie anschließend darzustellen.

2.2 Beschreibung der Programm-Phasen

HYBRID besteht aus drei Programm-Phasen, die hintereinander ausgeführt werden (Fig. 2).

2.2.1 USER-Phase

Die erste Phase enthält das vom Anwender geschriebene FORTRAN-Programm, das die Daten einliest bzw. erzeugt. Da das HYBRID-System sehr wenig Kernspeicherplatz belegt (siehe Anhang), können auch komplizierte Datenmanipulationen ausgeführt werden, die viel Kernspeicherplatz beanspruchen und deshalb bei konventionellen Programmen stets vorher ausgeführt werden müssen. Für oft wiederkehrende Operationen, wie Lorentztransformationen, Massenberechnungen usw. kann der Benutzer Unterprogramme verwenden (siehe Anhang), die automatisch dem Programm zugefügt werden. Die darzustellenden Daten werden durch Unterprogramm-Aufruf den entsprechenden Verteilungen zugewiesen. HYBRID schreibt die mit der Verteilungsnummer versehenen Werte (bzw. Wertepaare) als HYBRID-Records (Fig.3) auf einen Zwischendatensatz.

Um den Wünschen der verschiedenen Anwender gerecht zu werden, gibt es für den 1. JOBSTEP zwei unterschiedliche Arbeitsweisen:

- HYBRID ist wie ein Unterprogramm aufrufbar (Fig. 5a). -

Diese Arbeitsweise ist bequem für Anwender, die ihr Hauptprogramm schon fertig geschrieben haben und ihre Daten in geeigneter Weise graphisch darstellen wollen. Die Befehle zum Aufbau von Verteilungen sind bis auf Initialisierungsbefehle am Anfang und Ende des Benutzerprogramms dieselben wie bei der zweiten Arbeitsweise:

- HYBRID bildet das Rahmenprogramm (Fig. 5b) -

Der Benutzer schreibt ein oder mehrere Unterprogramme, die bestimmte Namen haben und an definierten Positionen im logischen Ablauf aufgerufen werden, wie z.B. Einlesen der Daten, Manipulation und Darstellung der Daten. HYBRID sorgt hierbei außerdem für einen geregelten Abbruch des Programms, bevor das System es bei Zeitüberschreitung gewaltsam zuende führt. Diese Arbeitsweise wird vorzugsweise für Daten angewendet, die aus Blasenkammer, Streamerkammer und ähnlichen Experimenten gewonnen sind.

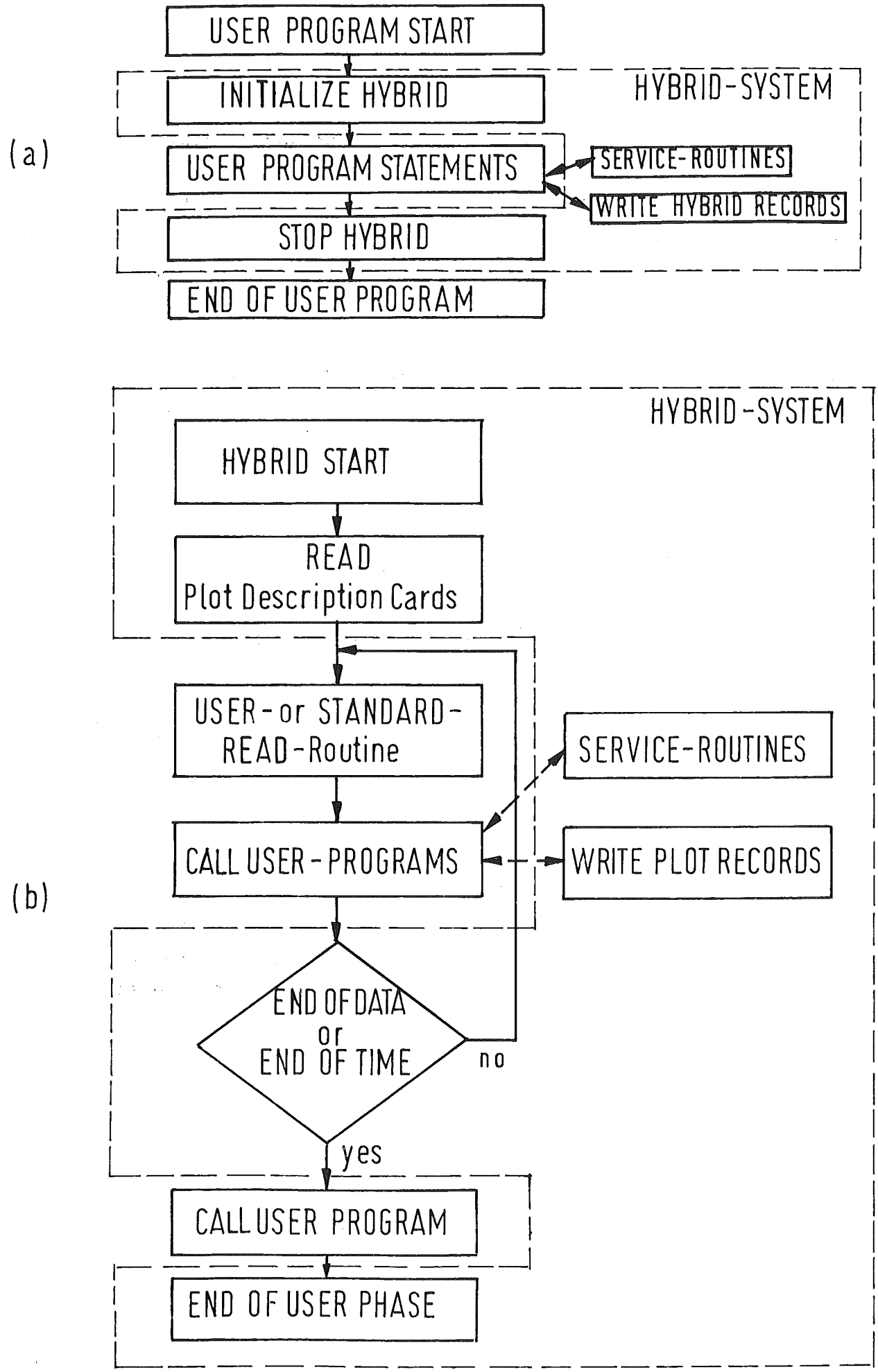

Fig. 5 Flußdiagramm der USER-PHASE

In vielen Fällen ist das Schreiben von Datenkarten, die dem Programm die Parameter für die gewünschten Verteilungen (Anfangswerte, Intervallbreiten) eingeben sollen, unbequem und zeitraubend. HYBRID kommt auf Wunsch auch ohne diese Karten aus. Allerdings ist die statistische Verteilung der Daten eine Voraussetzung für das vernünftige Wirken dieser Automatik[8]. Sie ist vor allem sehr bequem zur Darstellung von Daten, deren Wertebereich noch unbekannt ist.

Als Beispiel für die geringe Mühe, die der Benutzer mit HYBRID hat, diene der im Anhang aufgeführte Job.

2.2.2 SORT - Phase

Die SORT-Phase sortiert die Daten des Zwischenspeichers (z.B. Platte) nach Verteilungsnummern. Als Sortierprogramm sollte in jedem Fall ein Programm in Maschinensprache gewählt werden, um höchste Effektivität zu wahren. Zur Einsparung von REAL-TIME ist es wichtig, die Anzahl der Arbeits-Zwischenspeicher möglichst groß zu machen.

2.2.3 PRINT-Phase

In der PRINT-Phase wird der sortierte Zwischendatensatz bearbeitet. Die Daten werden in einem mit 120 x 120 x 4 Bytes eingerichteten Speicherplatz solange akkumuliert, bis die Verteilungsnummer im HYBRYD-Record wechselt. Dann wird die entsprechende Verteilung ausgedruckt, der Kernspeicherplatz wieder gelöscht und der sortierte Datensatz weiter gelesen.

---

[8] Zur Verfügung gestellt von V.Blobel

2.3 Zusammenfassung :

Die Vorteile von HYBRID gegenüber konventionellen Analyse-Programmen sind

- geringer Kernspeicherbedarf, dadurch Verlegung von Vorbereitungsarbeiten bzw. Datenerzeugung in den Analyse-JOB möglich
- unbegrenzte Anzahl von ein- und zweidimensionalen Verteilungen bei einmaliger Bearbeitung der Daten
- der Benutzer braucht nur einige FORTRAN-Befehle und keine neue Pseudo-Sprache zu lernen.

3. Anhang

HYBRID wird seit 5 Jahren bei DESY benutzt; zunächst auf einer IBM 7044 und in den letzten drei Jahren auf IBM/360 Modellen 75 bzw. 65. Im folgenden wird die Realisierung von HYBRID bei DESY näher beschrieben. HYBRID ist in dieser Form auf alle anderen IBM/360 (mit Kernspeicher >150K und genügender Anzahl externer Speicher) zu übertragen. Die Probleme, die bei der Implementation auf Nicht-IBM-Maschinen auftreten, werden weiter unten aufgezeigt.

3.1 HYBRID auf der IBM/360 bei DESY

HYBRID wurde als Prozedur in die PROCLIB der IBM/360 aufgenommen. Auf diese Weise muß der Benutzer nur ein Mindestmaß an Kontrollkarten schreiben und braucht sich um die Einrichtung von Zwischendatensätzen usw. nicht zu kümmern(s.a. 3.1.4).

3.1.1 USER-Phase

Die USER-Phase besteht aus 3 sog.JOBSTEP's (Fig. 6)

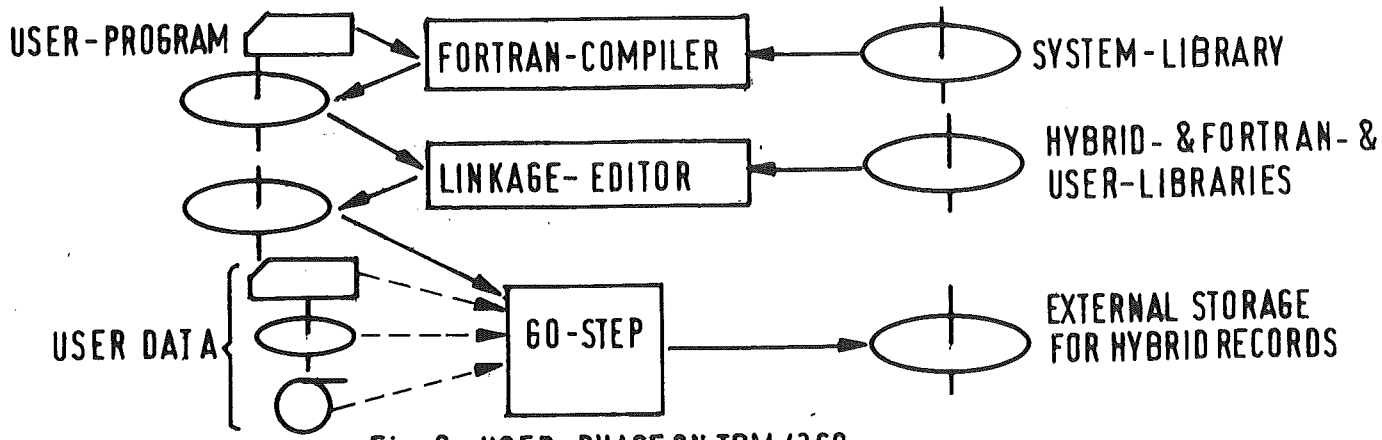

Fig.6 USER-PHASE ON IBM/360

Der erste STEP behandelt das vom Benutzer geschriebene FORTRAN-Programm. Dieses enthält die FORTRAN-Befehle, die z.B. zum Einlesen der Daten (für Standard-Datensätze gibt es fertige Unterprogramme) oder zur Erzeugung von Daten (bei Monte-Carlo-Programmen) dienen; weiterhin die Befehle zur Manipulation der Daten (Standardrechnungen werden mit vorhandenen Unterprogrammen ausgeführt) und die Zuweisung der Daten zu den gewünschten Verteilungen.
Dieses FORTRAN-Programm (s.a. Beispiel unten) wird vom Compiler übersetzt und auf eine Zwischenplatte geschrieben. Diese liest im 2. STEP der IBM-LINKAGE-EDITOR ein, fügt verlangte Unterprogramme aus den FORTRAN-, HYBRID- und USER-Libraries hinzu und bereitet das Programm für den nachfolgenden GO-STEP (3. STEP) vor. Der GO-STEP bearbeitet die Daten den Benutzerwünschen entsprechend und schreibt HYBRID-Records (Fig. 3) auf eine Zwischenplatte.
Für den GO-STEP werden vom HYBRID-System in der gegenwärtigen Fassung etwa 80k Bytes belegt. Der übrige Speicherplatz steht für Benutzerprogramme zur Verfügung.

Als Zwischenplatte für die HYBRID-RECORD's wird ein 2314-Magnetplattenstapel verwendet. Er faßt maximal $1.5*10^6$ Records. In der bisherigen Praxis zeigte es sich, daß in normalen Fällen stets ein halber 2314-Plattenstapel ausreichte.

In der Regel wird der im GO-STEP zur Verfügung stehende Kernspeicherplatz durch Benutzerprogramme nicht voll ausgenutzt.
Es wurde daher auf der IBM/360 eine dynamische Speicherplatzbelegung der vom Benutzer übriggelassenen Plätze eingeführt.
Auf diesen Plätzen werden so viel (eindimensionale) Verteilungen wie möglich gespeichert. Dadurch reduziert sich für diese Verteilungen die Anzahl der Ein- und Ausgabeoperationen ganz wesentlich.

In der HYBRID-Unterprogramm-Bibliothek befinden sich vier Klassen von Programmen:

A) HYBRID-SYSTEM-Programme

B) Unterprogramme zum Aufbau von Verteilungen:

1. Histogramme
2. Zweidimensionale Verteilungen
3. Ideogramme

und speziell für die Hochenergiephysik:

4. CHEW-LOW KONTUREN
5. DALITZ KONTUREN

6.Van Hove-Plots

Der CPU Zeitbedarf für einen Eintrag in eine Verteilung (= 1 HYBRID-RECORD) beträgt etwa 1 msec (IBM/360-75).

C) Programme zum Testen von Auswahlkriterien
Interpolationsprogramme

D) Spezielle Programme für die Hochenergiephysik
1. Einleseprogramme für Standard-Datensätze (GRIND, THRESH, WELAGA, usw.)
2. Rechnung mit Vierervektoren (Addition, Subtraktion, Lorentztransformation, Massenberechnungen, Zerfalls- und Erzeugungswinkel, ...)

3.1.2 SORT-PHASE

Die SORT-PHASE bedient sich des IBM-SORT-Programms mit der sog. BALANCED-SORT-Technik. Um bestmögliche REAL-TIME-Zeiten zu erzielen, werden sechs 2314-Einheiten als Arbeitsplatten verwendet (Fig. 7). Als Sortierbegriff werden die ersten 32 bit des HYBRID-RECORDs, also die Verteilungsnummern verwendet.

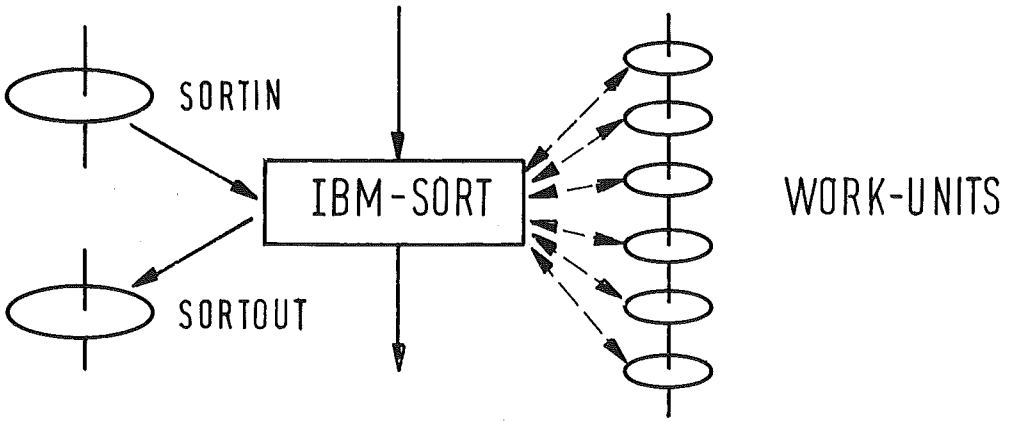

Fig.7 SORT-PHASE

In der Fig.8 wird der bei dieser Konfiguration nötige CPU-TIME-Bedarf dargestellt.

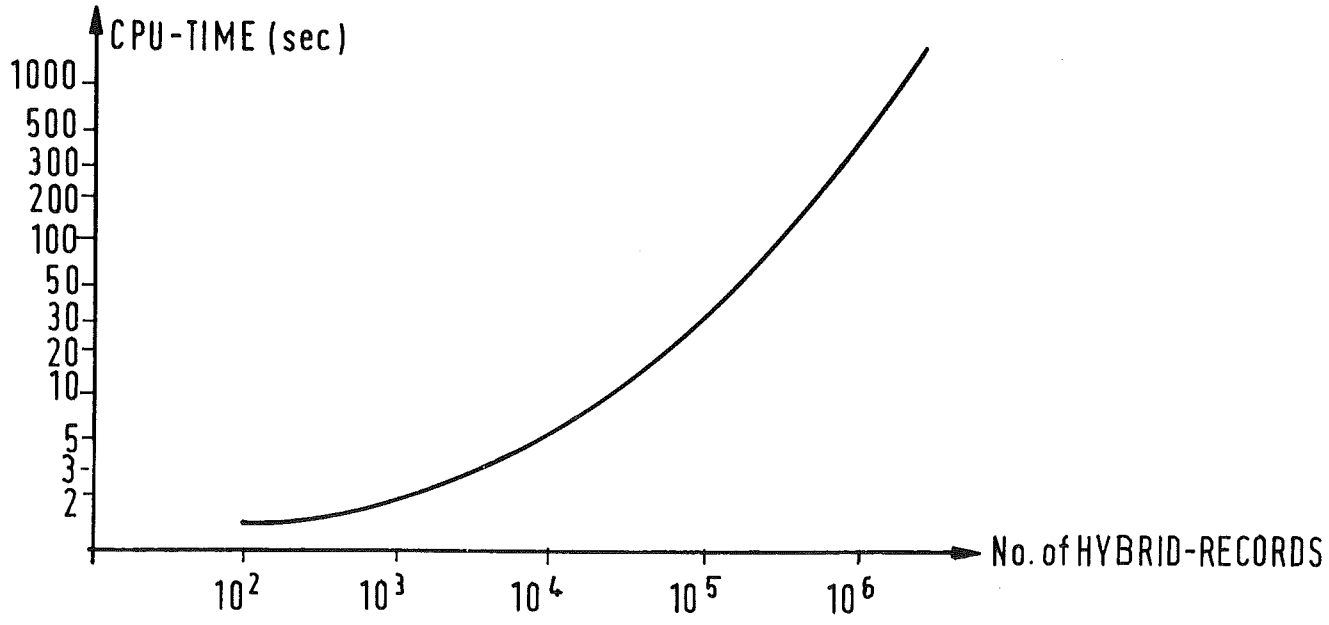

Fig. 8 CPU-TIME Bedarf der SORT-PHASE

Der Bedarf an REAL-TIME ist sehr von der Maschinenkonfiguration und dem SYSTEM-Modus abhängig (z.B. TIME-SHARING, MFT, MVT). Bei DESY (MVT-Betrieb) ist die REAL-TIME etwa um den Faktor 1.5-2 größer als die CPU-TIME.

3.1.3 PRINT-PHASE

Die PRINT-PHASE bearbeitet den nach Verteilungsnummern sortierten Zwischendatensatz. Das PRINT-Programm ist in FORTRAN programmiert und benötigt etwa 150k Bytes.

Die für die PRINT-PHASE benötigte CPU-TIME hängt einmal von der Anzahl der Verteilungen ab, zum zweiten von der Anzahl der HYBRID-Records. Bei einer Anzahl von 100 Verteilungen und 10^5 HYBRID-Records, bei einer Gleichverteilung der Einträge in die Verteilungen, ergeben sich etwa 0,5 msec pro HYBRID-Record.

3.1.4 BEISPIEL

Die Einfachheit der HYBRID-Programmierung zeigt der folgende Job: Es werden 500 zweidimensionale Verteilungen (DPLOT) von Daten angefertigt, die auf einem Datensatz vorhanden sind:

```
//PLOTJOB JOB 'HYBRID-PLOTS',ANYNAME
// EXEC HYBRID,TIME=10
      DIMENSION X(500), Y(500)
      CALL INIT
   10 READ (1,END=100) X,Y
      DO 20 I=1,500
   20 CALL DPLOT (Y(I), X(I), I)
      GOTO 10
  100 CALL END
      STOP
      END
//GO1.FT01F001 DD ... Datendefinition ...
```

Datenkarten sind nicht nötig, da das Programm in diesem Fall Anfangswerte und Intervallbreiten nach statistischen Methoden berechnet.

3.2 Implementation von HYBRID bei anderen Systemen

HYBRID ist in IBM-FORTRAN IV Level H programmiert. Erforderlich ist eine Wortlänge von 32 Bits. Ganz spezielle Aufgaben (z.B. dynamische Belegung von Speicherplatz) werden von Programmen in Maschinensprache wahrgenommen, jedoch sind diese zur prinzipiellen Funktion von HYBRID nicht unbedingt notwendig.

Als Problem stellt sich bei anderen Rechenmaschinen meist das Fehlen von Sortierprogrammen und der zugehörigen Hardware heraus. Sortierprogramme sind jedoch (zwar nicht ganz so effektiv wie Maschinenprogramme) auch in FORTRAN programmierbar, und als externe Speichermedien sind selbst Magnetbänder ziemlich effektiv.

Der erste Teil von HYBRID beansprucht mit etwa 80k Bytes (durch Verzicht auf einigen Komfort auf 60k reduzierbar) weniger als andere konventionelle Analyseprogramme, da man den Platz für zweidimensionale Verteilungen einspart.

Das Ausdruckprogramm ist auf maximale Geschwindigkeit ausgerichtet und benötigt etwa 150k Bytes.

Zu dem gesamten Speicherplatzbedarf kommen noch stets die Puffer für die Zwischendatensätze hinzu; diese sind stark von der Maschinenkonfiguration abhängig.